

THE REVIEWER'S GUIDE TO
QUANTITATIVE METHODS
IN THE SOCIAL SCIENCES

SECOND EDITION

revise
accept
reject

Edited by **Gregory R. Hancock,**
Laura M. Stapleton, and Ralph O. Mueller



The Reviewer's Guide to Quantitative Methods in the Social Sciences

The Reviewer's Guide to Quantitative Methods in the Social Sciences provides evaluators of research manuscripts and proposals in the social and behavioral sciences with the resources they need to read, understand, and assess quantitative work. 35 uniquely structured chapters cover both traditional and emerging methods of quantitative data analysis, which neither junior nor veteran reviewers can be expected to know in detail. The second edition of this valuable resource updates readers on each technique's key principles, appropriate usage, underlying assumptions and limitations, providing reviewers with the information they need to offer constructive commentary on works they evaluate. Written by methodological and applied scholars, this volume is also an indispensable author's reference for preparing sound research manuscripts and proposals.

Gregory R. Hancock is Professor, Distinguished Scholar-Teacher, and Director of the Measurement, Statistics and Evaluation program in the Department of Human Development and Quantitative Methodology at the University of Maryland, as well as Director of their Center for Integrated Latent Variable Research.

Laura M. Stapleton is Professor in the Measurement, Statistics and Evaluation program in the Department of Human Development and Quantitative Methodology at the University of Maryland.

Ralph O. Mueller is Vice Chancellor for Academic Affairs and Provost at Purdue University Northwest.

Updated and even more useful, this much-needed volume fills a gap for review consultation and instructional purposes. Highly recommended!

Michael G. Vaughn, Saint Louis University

The first edition of this book belongs on every reviewer's bookshelf. The second edition is even better and covers topics missed in the first.

David L. Streiner, McMaster University

As an editor for more than 20 years, I had long wrestled with what graduate students and reviewers need to understand and address when evaluating the quality of the quantitative analyses reported in manuscripts. That problem is made even more frustrating by the range of quantitative methods populating the educational research literature. Thanks to this outstanding volume those nagging concerns have largely been put to rest. What these respected editors have compiled are 35 insightful chapters, each devoted to a particular quantitative method and written by an acknowledged expert. Each chapter not only succinctly overviews a given technique, but also delineates the musts and shoulds of its reporting, which are summarized in an easily referenced table. I plan to make this essential guide required reading for all my graduate students and for every editorial board member.

**Patricia A. Alexander, Jean Mullan Professor of Literacy, Distinguished Scholar-Teacher,
University of Maryland. Senior Editor, *Contemporary Educational Psychology***

Greg Hancock and his colleagues have done it again. The second edition of the Reviewer's Guide to Quantitative Methods in the Social Sciences offers 35 chapters written by top-of-the line quantitative researchers who inspire, instruct and illuminate. The chapters provide key information that is essential in evaluating a wide swath of methods including basic and multivariate statistics, research design, statistical inference procedures, psychometrics, latent variable methods, modeling and more. Every social scientist would benefit from this gem of a volume that cannot help but leave readers more informed, enlightened, and empowered

Lisa L. Harlow, Professor of Psychology, University of Rhode Island

The Reviewer's Guide to Quantitative Methods in the Social Sciences, Second Edition is an essential resource for editors, reviewers, researchers, and graduate students who desire to produce and disseminate accurate and meaningful quantitative research. Each of the 35 chapters provides a comprehensive overview of a particular aspect of quantitative study design and/or analytic technique, indicating where (within a research report) and how critical information should be conveyed. Of particular value are the recommendations for appropriate language to use in describing and interpreting quantitative findings, including admonitions against common reporting fallacies, gaps, and other improprieties. It goes without saying that all editors and reviewers should consult this authoritative guide as a de facto set of rigorous standards for evaluating quantitative inquiry, and emerging researchers would benefit tremendously from working with these insightful chapters as they develop a fundamental understanding of designing, conducting, and reporting their work.

John Norris, Senior Research Director, Educational Testing Service

The Reviewer's Guide to Quantitative Methods in the Social Sciences

Second Edition

Edited by

**Gregory R. Hancock, Laura M. Stapleton,
and Ralph O. Mueller**

Second edition published 2019
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2019 Taylor & Francis

The right of Gregory R. Hancock, Laura M. Stapleton, and Ralph O. Mueller to be identified as the authors of the editorial material, and of the authors for their individual chapters, has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilized in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

First edition published by Routledge 2010

Library of Congress Cataloging-in-Publication Data

Names: Hancock, Gregory R., editor. | Stapleton, Laura M., editor. |
Mueller, Ralph O., editor.

Title: The reviewer's guide to quantitative methods in the social sciences /
edited by Gregory R. Hancock, Laura M. Stapleton, and Ralph O.
Mueller.

Description: Second Edition. | New York : Routledge, 2019. | Revised
edition of The reviewer's guide to quantitative methods in the social
sciences, 2010. | Includes bibliographical references and index.

Identifiers: LCCN 2018029418 (print) | LCCN 2018030586 (ebook) |
ISBN 9781315755649 (e-book) | ISBN 9781138800120 (hardback) |
ISBN 9781138800137 (pbk.) | ISBN 9781315755649 (ebk)

Subjects: LCSH: Social sciences—Research—Methodology. | Social
sciences—Statistical methods.

Classification: LCC H62 (ebook) | LCC H62 .R466 2019 (print) | DDC
300.72/1—dc23

LC record available at <https://lccn.loc.gov/2018029418>

ISBN: 978-1-138-80012-0 (hbk)

ISBN: 978-1-138-80013-7 (pbk)

ISBN: 978-1-315-75564-9 (ebk)

Typeset in Minion
by Swales & Willis Ltd, Exeter, Devon, UK

Dedicated to the unheralded reviewers whose words frustrate and challenge us—and make our work stronger for it.



Taylor & Francis
Taylor & Francis Group
<http://taylorandfrancis.com>

Contents

<i>Preface</i>	x
1 Analysis of Variance: Between-Groups Designs ROBERT A. CRIBBIE AND ALAN J. KLOCKARS	1
2 Analysis of Variance: Repeated-Measures Designs LISA M. LIX AND H. J. KESELMAN	15
3 Canonical Correlation Analysis XITAO FAN AND TIMOTHY R. KONOLD	29
4 Cluster Analysis DENA A. PASTOR AND MONICA K. ERBACHER	42
5 Correlation and Other Measures of Association JILL L. ADELSON, JASON W. OSBORNE, AND BRITTANY F. CRAWFORD	55
6 Effect Sizes and Confidence Intervals FIONA FIDLER AND GEOFF CUMMING	72
7 Event History and Survival Analysis PAUL D. ALLISON	86
8 Factor Analysis: Exploratory and Confirmatory DEBORAH L. BANDALOS AND SARA J. FINNEY	98
9 Generalizability Theory AMY HENDRICKSON AND PING YIN	123
10 Interrater Reliability and Agreement WILLIAM T. HOYT	132
11 Item Response Theory and Rasch Modeling R. J. DE AYALA	145
12 Latent Class Analysis KAREN M. SAMUELSEN AND C. MITCHELL DAYTON	164
13 Latent Growth Curve Models KRISTOPHER J. PREACHER	178

14	Latent Transition Analysis DAVID RINDSKOPF	193
15	Latent Variable Mixture Models GITTA LUBKE	202
16	Logistic Regression and Extensions ANN A. O'CONNELL AND K. RIVET AMICO	214
17	Log-Linear Analysis RONALD C. SERLIN AND MICHAEL A. SEAMAN	235
18	Mediation and Moderation PAUL E. JOSE	248
19	Meta-analysis S. NATASHA BERETVAS	260
20	Monte Carlo Simulation Methods DANIEL MCNEISH, STEPHANIE LANE, AND PATRICK CURRAN	269
21	Multidimensional Scaling CODY S. DING AND SE-KANG KIM	277
22	Multilevel Modeling D. BETSY MCCOACH	292
23	Multiple Regression KEN KELLEY AND SCOTT E. MAXWELL	313
24	Multitrait–Multimethod Analysis KEITH F. WIDAMAN	331
25	Multivariate Analysis of Variance KEENAN A. PITUCH AND WANCHEN CHANG	348
26	Nonparametric Statistics MICHAEL A. SEAMAN	362
27	Power Analysis KEVIN R. MURPHY	380
28	Propensity Scores and Matching Methods ELIZABETH A. STUART	388
29	Reliability and Validity RALPH O. MUELLER AND THOMAS R. KNAPP	397

30	Research Design	402
	SHARON ANDERSON DANNELS	
31	Single-Subject Design and Analysis	417
	ANDREW L. EGEL, CHRISTINE H. BARTHOLD, JENNIFER L. KOUO, AND FAYEZ S. MAAJEENY	
32	Social Network Analysis	434
	TRACY SWEET	
33	Structural Equation Modeling	445
	RALPH O. MUELLER AND GREGORY R. HANCOCK	
34	Structural Equation Modeling: Multisample Covariance and Mean Structures	457
	RICHARD G. LOMAX	
35	Survey Sampling, Administration, and Analysis	467
	LAURA M. STAPLETON	
	<i>List of Contributors</i>	482
	<i>Index</i>	492

Preface

Volume Background

A cornerstone of research institutions and agencies around the world is the creation of new knowledge that often is generated through utilizing quantitative research methods. The dissemination of such knowledge through specialized journals, application-oriented outlets, and technical reports is typically filtered through a rigorous peer review process to ensure work of the highest possible quality. Reviewers, and the editors they serve, thus operate in the critical role of gatekeeper and must, collectively, be held accountable for the integrity of the research enterprise.

The quantitative skills that reviewers bring to this role tend to fall into two categories: expertise in methods they use fairly regularly and competently in their own research, and knowledge from academic or professional training that has largely laid dormant since that initial exposure. This limiting state of affairs, which is exacerbated as cohorts of new researchers are trained in increasingly advanced data analysis techniques, can force quantitatively uninitiated reviewers to confine their critical commentary primarily to the content-area portions of a manuscript. In the end, these reviewers are operating by assuming that someone else is tending the methodological gate, and editors are left in the difficult position of burdening those few reviewers who are quantitatively current while constantly having to update their stable of adjunct reviewers with highly specific areas of methodological expertise.

In all fairness, reviewers, whether novice or veteran, should not be expected to have a command of all data analysis methods used in modern research. We believe that they should, however, maintain some broad and evolving awareness of methodological developments, and make an honest attempt to evaluate the analytical methods that are at the core of a manuscript's intellectual contribution. The current volume was born out of a desire to assist reviewers in meeting this professional responsibility. In particular, it is designed as a reference tool specifically for reviewers of manuscripts and proposals in the social sciences and beyond, addressing a broad range of traditional and emerging quantitative techniques and providing easy access to critical information that authors should have considered and addressed in their submissions.

Volume and Chapter Structure

This second edition has 35 chapters arranged alphabetically by title, with each chapter addressing a particular quantitative method or area. These include the sound practice of research (e.g., research design, survey sampling, power analysis, propensity score and matching methods), the general linear model (e.g., analysis of variance, multiple regression, hierarchical linear modeling), the generalized linear model (e.g., logistic regression, log-linear analysis), measurement (e.g., item response theory, generalizability theory, multidimensional scaling), and latent structure methods (e.g., latent class analysis, latent growth curve models, structural equation modeling). Thirty chapters from the first edition have been updated; one original chapter was eliminated with its content now touched upon in other chapters' revised content, and *five new chapters have been added to the second edition: "Mediation and Moderation," "Monte Carlo Simulation Methods," "Nonparametric Statistics," "Propensity Scores and Matching Methods," and "Social Network Analysis."* As with the first edition, the structure across all chapters is the same, consisting of the following three sections:

- Method overview
- Table of desiderata
- Explications of desiderata.

Method overview. Each chapter starts with a brief introduction and overview of the method(s) being addressed in the chapter. This is not meant to teach the reader the topic area per se, but rather to provide an orientation and possible refresher. This section concludes with useful references for the reader who wants more introductory and/or advanced information on that chapter's topic.

Table of desiderata. After each chapter's method overview, a numbered list is provided of key elements (desiderata) that should be addressed in any study using that chapter's method(s). This table serves to provide essential evaluation criteria that a reviewer should consider when judging a manuscript's methodological approach to data analysis, including the technique's key principles, appropriate usage, underlying assumptions, and limitations. For each desideratum the section(s) of a manuscript in which the specific issue should most likely be addressed is denoted with abbreviations: I for Introduction, M for Methods, R for Results, and D for Discussion.

Typically, the desiderata appear in a single table; in some cases, the table is partitioned by special applications of a certain method (e.g., Chapter 21, "Multidimensional Scaling"). In a couple of cases, the desiderata are presented in two separate tables due to the bifurcated nature of the topic (e.g., Chapter 8, "Factor Analysis," with tables for exploratory and confirmatory methods). The user of this volume will also note that there are many desiderata in common across chapters, including such elements as making connections between research questions and the analytic method at hand, explicitly addressing how missing data and outliers were handled, reporting the software and version used, and so forth. Although these could have been culled for a separate table at the beginning of the volume, we believe that having them contained within each chapter as they pertain to the specific method is in keeping with the reference nature of this guide. In this manner, a chapter's table(s) of desiderata may be used by a reviewer as a checklist to evaluate a manuscript's methodological soundness.

Explications of desiderata. Following each chapter's table(s) of desiderata are corresponding explications for each numbered desideratum. For a reader already thoroughly familiar with a particular desideratum, a given explication may be unnecessary; we expect, however, that most readers will benefit from the supporting explanation, elaboration, and any additional references specific to that desideratum. For example, if a desideratum calls for a manuscript to explicitly examine the assumptions underlying a particular analytical technique, the explication might provide a treatment of the inferential consequences of failing to examine those assumptions as well as the preferred methods for conducting and presenting the results of such an examination. The explications of the desiderata constitute the main body of each chapter, justifying each desideratum in light of accepted practice and the supporting methodological literature.

The Past, Present, and Future of This Volume

Any time someone offers recommendations as to proper practice or conduct, reasonable people will disagree. The current volume, methods, and desiderata are no different. In fact, we know that some knowledgeable readers will take issue with particular chapter authors' portrayal of a given method's best practice. That said, in the preparation of this second edition, as with the first edition, we have encouraged chapter authors to try to convey what each technique's methodological literature considers to be the most accepted practices associated with a given method. Further, we understand that even if there is currently agreement on a technique's best practices, the growing methodological literature might determine otherwise in the future.

As stated in the preface to the first edition, we view this volume as a living resource. As quantitative methodologies evolve, not just in preferred practice of existing techniques but also in the development of new techniques, each edition of this guide will adapt as well, as reflected in the updates and expansions contained in this second edition. Changes in the current volume have been in large part due to the positive response to the first edition from manuscript reviewers and journal editors, from grant proposal panelists, and from the researchers themselves. We continue to encourage readers' correspondence with the chapter authors and the editors as part of the on-going dialog about specific methods and their desiderata, thus helping to keep this volume responsive to changes in appropriate practice. We look forward to this guide serving as an ongoing and dynamic resource for reviewers, as well as for the authors themselves (e.g., applied researchers, university faculty, and graduate students) when designing their own research projects.

Gregory R. Hancock, Laura M. Stapleton, and Ralph O. Mueller
June 29, 2018

1

Analysis of Variance *Between-Groups Designs*

Robert A. Cribbie and Alan J. Klockars

Between-groups analysis of variance (ANOVA) is one of the most commonly used techniques to analyze data when the intent of the research is to determine whether one or more categorical independent variables (IVs) relates to a continuous dependent variable (DV). Causal statements regarding this relation are often clearest with random assignment of subjects to the various treatment groups (although see Chapter 30 for dealing with quasi-experimental designs). The broad area of ANOVA consists of significance tests to determine if a non-random relation exists between IV(s) and the DV, follow-up tests to investigate more thoroughly the nature of such a relation, and measures of the strength of the relation.

The following is an overview of some of the types of designs and experiments that can be analyzed by *between-groups ANOVA*. The number of IVs will determine the type of design. For example, a single IV design is called a *one-way design*. There are also ANOVA designs that have a number of different IVs. If, for example, there were three IVs that were completely crossed with one another, it would be called a *three-way factorial design*. Crossing IVs in a *factorial experiment* typically enriches the major theory being investigated, providing for additional main effects as well as interactions between the factors.

A *random factor* is one that has the levels of the IV randomly chosen from some universe of possible levels (e.g., dosages of a medication, hours of instruction), which is distinguishable from a *fixed factor*, where all levels of the IV that are of interest are included (e.g., treatment and control). Random factors may be crossed with the fixed factors yielding a *mixed model*. If one IV is nested within another, the design can be labeled *hierarchical*. Individual difference variables might also be incorporated into ANOVA to reduce nuisance variability, as in a *randomized block design*, or as a continuous control variable within an *analysis of covariance* (ANCOVA). There is a rich literature in ANOVA, including texts written for experimental design by Kirk (1995), Maxwell and Delaney (2003), and Keppel and Wickens (2004).

1. Dependent Variables

Researchers must clearly outline the dependent variable under study, most importantly describing in detail how that outcome is being measured (for designs with multiple outcomes, see Chapter 24, this volume). The outcome variable is the set of numbers that we have collected that are used as a

Table 1.1 Desiderata for Analysis of Variance, Between-Groups Designs.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. The dependent variable(s) under study are outlined with a discussion of their importance within the field of study.	I
2. Each discrete-level independent variable is defined and its hypothesized relation with the dependent variable is explained.	I
3. A rationale is provided for the simultaneous inclusion of two or more independent variables and any interaction effects are discussed in terms of their relation with the dependent variable.	I
4. Appropriate analyses are adopted when the research hypothesis relates to the equivalence of means.	I
5. The inclusion of any covariate is justified in terms of its purpose within the analysis.	I
6. The research design is explained in detail, including the nature/measurement of all independent and dependent variables.	M
7. In randomized block designs, the number, nature, and method of creation of the blocks is discussed.	M
8. The use of a random factor is justified given the hypotheses.	M
9. In hierarchical designs, the rationale for nesting is explained and the analysis acknowledges the dependence in the data.	M
10. In incomplete designs, or complex variants of other designs, sufficient information and references are provided.	M
11. A rationale is given for the number of participants, the source of the participants, and any inclusion/exclusion criteria used.	M
12. Missing data and statistical assumptions of the model are investigated and robust methods are adopted when issues arise.	M, R
13. The final model is discussed, including defining and justifying the chosen error term and significance level.	M, R
14. Follow-up strategies for significant main effects or interactions are discussed.	M, R
15. Effect size and confidence interval information are provided to supplement the results of the statistical significance tests.	R
16. Appropriate language, relative to the meaning and generalizability of the findings, is used.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

proxy for the theoretical dependent variable. The authors must defend the outcome variable that is being used by, for example, establishing sufficient construct validity within the population being studied. The reader should be able to see the obvious fit between the measure and the construct.

For ANOVA, the outcome measure should be measured on a continuous scale. All analyses and decisions made in tests of significance are about the characteristics of the outcome measure, not directly on the theoretical dependent variable. If the means for groups differ statistically significantly on the outcome variable, this strongly suggests a relation between the grouping variable and the outcome. The researcher, however, does not typically want to discuss differences on the unique outcome measure used but rather differences on the theoretical dependent variable that the test is supposed to tap. Only to the extent that the outcome measure provides a valid measurement of the theoretical construct will the conclusions drawn be relevant to the underlying theory.

The overall mean and variance of the outcome measure must be such that potential differences among the groups can be detected. Outcomes where subjects have scores clustered near the minimum or maximum possible on the instrument (i.e., floor or ceiling effects) provide little opportunity to observe a difference. This will reduce power and obscure differences that might really be present. Floor or ceiling effects also typically result in distributions that are not normal in shape (see Desideratum 12).

Authors often employ measures used in previously published literature. Of concern is the appropriateness of the measure in the new setting, such as with different age or ability levels. In factorial experiments, particularly with randomized blocks designs where one of the factors is an individual difference measure on which subjects have been stratified, the inappropriateness of the measure at some levels of the blocking variable might incorrectly appear as an interaction of the blocks with the treatments. For example, consider a learning experiment where the impact of the treatment was measured by an achievement test and the subjects were blocked on ability. If it were true that one treatment was generally superior to the others across all levels of ability, the appropriate finding would be a main effect for treatment. However, if the outcome measure were such a difficult test that subjects in all but the highest ability group obtained essentially chance scores (which would not show treatment effects), while those in the highest ability group showed the true treatment effect, the effect would be declared an aptitude-treatment interaction.

2. Independent Variables

A between-groups ANOVA requires unique, mutually exclusive groups. Typically, the groups reflect: (1) fixed treatments, (2) levels of an individual difference variable, or (3) levels of a random variable (i.e., a random factor). Fixed treatment groups are created such that they differ in some aspect of the way the participants are treated. The differences among treatments capture specific differences of interest and thus are usually treated as a fixed factor with either qualitatively different treatments or treatments having different levels of intensity of some ordered variable. Categorical individual difference variables (e.g., race) may be included as the primary IV of interest or as blocking variables in a randomized block design. Random factors, where the levels of the IV are a subset of all potential levels available, must be treated different than fixed factors given that there is variability associated with the selection of levels.

Regardless of the nature of the IV, it is important that researchers explain the hypothesized relation between the IV and DV. Unless the researcher clearly outlines the study as exploratory, the nature of the relation between the IV and DV should be explicit. For example, in most cases it is not sufficient for a researcher to simply state that they are comparing the levels of the IV on the DV, but instead should justify why the IV was included in the study and specifically how they expect the levels of the IV to differ on the DV.

When treatment groups are created, the groups are the operational definitions of the theoretical independent variable of interest. In the report of the research, the relations described are generally in terms of the theoretical construct the groups were created to capture (e.g., stress) not in terms of the operations involved (e.g., group was told the assignment counted for 50% of their grade). The way the treatments are defined must clearly capture the essence of the theoretical variable. Authors must defend the unique characteristics of the treatment groups, clearly indicating the crucial differences desired to address the theoretical variable of interest. The treatments must have face validity in that the reader sees the obvious linkage between the theoretical variable of interest and the operations used to create that variable.

There are a number of common shortcomings in the construction of treatment groups that can result in low power or confounded interpretations. For example, differences in the wording of

reading prompts might have a subtle effect too small to be detected by an experiment unless an extremely large sample size is used. Low power can be the result of treatments that were implemented for too short a period or with insufficient intensity to be detected. Confounded effects can happen when groups differ in multiple ways, any one of which might produce observed differences. This can happen inadvertently, such as if one type of prompt required the study period to be longer than any of the other types of prompts. Any difference found might be due to either the differences in prompts or the differences in study time. In other experiments, the intent is to compare treatment groups that differ on complex combinations of many differences resulting in uncertainty regarding the ‘active ingredient’ that actually produced any group differences found.

With regard to individual difference variables, it is important that the researcher explains whether the variable is being used as a primary IV or as a nuisance factor to reduce random variability, and in either case provide a theoretical rationale for including the variable. With random factors the author must defend the choices made with particular attention to the extreme levels (see Desideratum 8).

3. Inclusion of Two or More Independent Variables

Many settings use factorial experiments where multiple IVs are crossed to facilitate the assessment of both main effects and interactions. Sometimes one IV is central to an investigator’s research program, with the other IVs included for control and/or to expand upon the theory being explored. Authors should defend the inclusion of all IVs relative to their relation to the central theory they are testing. The explanation of an IV’s role should include both the main and interaction effects anticipated. It is not necessary for the researcher to include the interaction in the model if no hypothesis concerning the interaction exists.

Fixed IVs generally enrich theory. The inclusion of an IV that is a random factor is often meant to show the generalizability of the central findings across levels of the random factor. The choice of the random factor should be justified relative to the need for greater generalizability concerning the main effects (see Desideratum 8).

Factorial designs provide more information than a single factor design, but the authors should recognize the costs involved in terms of complexity of interpretation. The more complex the design, the more difficult it is to understand what, if anything, really differs. Interactions alter the interpretation of the main effects and thus, if an interaction exists, the main effects should not be discussed. For example, imagine that the effect of test type (short answer versus multiple choice) on grades differs across males and females (i.e., a test type by sex interaction). In this situation, it is not appropriate to discuss the overall effect of test type because the nature of the relation depends on the sex of the student. If an interaction is not statistically significant, authors sometimes over-generalize main effects ignoring the possibility that the statistically non-significant interaction may be due to lack of power (i.e., a Type II error). Lack of interaction tests can be used if the primary hypothesis relates to negligible interaction or if there is a desire to remove inconsequential interaction terms from a model. Lack of interaction tests fall under the category of equivalence tests (see Desideratum 4).

Factorial designs with unequal numbers of observations across the cells require special attention as the main effects and interactions are not orthogonal (see Desideratum 12).

4. Difference-Based or Equivalence-Based Hypotheses

Often, researchers propose that a relation exists between the IV(s) and the DV but in other instances researchers propose that there is no relation between the IV(s) and the DV. Imagine that a researcher was interested in demonstrating that two treatments were equally effective, or that the effect of a treatment was similar across genders (i.e., a lack of interaction). In this instance, typical

ANOVA procedures are not appropriate. Alternatives to the one-way and factorial between-subjects ANOVA for testing a lack of relation are available (field of equivalence testing) and should be adopted in these situations. Further, follow-up tests for nonsignificant omnibus tests of equivalence require specific procedures that differ in important ways from those discussed in Desideratum 14.

5. Covariates

Authors should justify the inclusion of any covariate(s) relative to the theory being tested. Covariates can increase the power of the statistical tests within the ANOVA, but can also increase the Type I error rate if their selection allows for the capitalization on chance. A posteriori or haphazardly selected covariates can capitalize on random variability and should not be used. Covariate scores must not be a function of the treatment group and, in randomized designs, should be available before random assignment takes place. Designs in which the covariate is obtained at the same time as the outcome have the likelihood that the covariate scores will be altered due to treatment. A primary function of the covariate is to provide a way to statistically equate groups to which the treatments are applied (e.g., subjects are randomly assigned to treatments although the balance of characteristics is less than perfect); this function would be invalidated if the treatment changed the covariate score used. When individual difference IVs are included, covariates are often used to equate the groups on relevant variables. For example, if we wanted to compare smokers and non-smokers on a DV of interest, it would be important to control for variables that predict smoking (e.g., low self-esteem).

The use of ANCOVA adds several assumptions to the analysis that are not germane to an ANOVA (see Desideratum 12). Some, such as the covariate being a fixed factor and measured without error, are likely to be violated, but that is true of almost every use of ANCOVA. Thus, there is little value in requiring every research paper using a covariate to include a discussion of these. However, authors should have considered and discussed any likely violation of the assumptions of linear relation between the covariate and the outcome measure, homogeneous residual variances within treatments, and homogeneity of regression coefficients within treatments.

A violation of the assumption of homogeneous regression coefficients is itself a major finding analogous to an interaction between the covariate and the treatment. The finding of differences in slopes should be considered an important discovery that could potentially be of interest in future research. In a similar manner, the finding that there are differences in the variability of outcome scores about the regression lines for different treatments may be a finding worthy of reporting and discussion.

6. Research Design

The Methods section should clearly describe the characteristics of the experiment so that the reader can easily evaluate the appropriateness of the Results and Discussion sections relative to the actual experiment conducted. The Methods section should include a summary of the research design that allows the reader to review all of the critical elements of the experiment. The information should be sufficiently complete so that the reader can anticipate the sources of variation in the dependent variable(s) and the degrees of freedom for each source.

The characteristics related to the design should include the following elements. The authors should indicate the specific nature of the research design (e.g., the research was a true experimental design with random assignment of subjects or the research used an individual difference variable). The number of factors and the number and nature of levels of each factor should be stated, as well as the resulting total number of treatment groups. Each factor should be identified as either a fixed

or a random factor within the experiment. If the experiment includes random factors, the author should clearly indicate the appropriate error term for the treatment effects and interactions that are impacted by the presence of the random factor (see Desideratum 8).

If appropriate, the authors should state that the factors are completely crossed, although the readers will assume a completely crossed experiment unless specifically informed that a factor was either incompletely crossed (as in a lattice design) or was nested. In some multifactor experiments, there are treatment groups that do not completely fit within the factorial structure. For example, if a reading study varied the types of prompts given to students, a second factor might be whether the prompts were given immediately prior to reading the passage or immediately after. A control group having no prompts could not incorporate this placement factor since the factor presumes the presence of some sort of prompt. The presence of this type of treatment group should be carefully acknowledged and treated in an appropriate manner. The author should also provide any technical term that might help summarize the experimental design (e.g., Latin square; see Desideratum 10).

7. Randomized Block Designs

The term *randomized block* will be used to describe the case where an individual difference variable is used to create levels (blocks) of a factor and where participants are then randomly assigned to treatments from within the blocks. The rationale for including the individual difference variable should be clearly stated. This rationale should distinguish between a motive of including the individual difference variable to remove variability in scores that would otherwise be considered random, and thus increase the power of the main effect, and a motive of exploring the relation between the individual difference variable and the treatment variable via a test of the interaction.

The particular measure of the individual difference variable used to create blocks should also be justified as a valid measure in the experimental setting. When the motive for including the individual difference variable is to explore the way that the treatments provide differing effects depending on the level of the blocks (i.e., an interaction effect), the rationale should be clearly outlined. The measure of the individual difference variable must be a valid measure of the theoretical construct in order to make meaningful interpretations of the interaction. This is less of a concern when reduction of the error variance is the sole motive for including the individual difference measure as there are no theoretically grounded tests that depend on the meaning of the individual difference variable.

The number of blocks used should be justified relative to the motive for including the individual difference variable, the strength of the relation between the individual difference variable and the outcome measure, and the quality of the individual difference measure. The method of creating blocks should be explained with particular attention to whether the blocks were independently defined followed by the recruitment of participants who fit into those blocks, or were blocks created post hoc out of a set of n available participants. If independently defined levels of the blocking measure were used, the authors should explain how (for continuous variables) the points used to define the levels were determined, with particular concern for the blocks on both ends of the continuum. In other words, the cut points affect the level of control of the blocking variable and thus should be meaningful.

The blocking variable can be considered either random or fixed, depending on how it was created and the extent to which the results are to be generalized. This distinction is important in determining whether the subjects-within-treatments (for fixed) or the block-by-treatment interaction (for random) should be used as the appropriate error term. If subjects-within-treatments is used as the error term, the generalizations should be carefully limited to the specific levels of the blocking

variable created or observed. However, the block-by-treatment interaction as an error term allows for greater generalization across the universe of possible blocks.

8. Random Factors

When a random factor is included, its use and the selection of the levels utilized from the pool of all potential levels available should be justified. The term *mixed model* is commonly used to describe a factorial experiment in which at least one of the factors is a random variable and at least one of the others is fixed. The primary purpose of a random factor is to argue for the generalizability of the fixed effect across the levels of the random factor, rather than just over the specific levels sampled.

Some random factors, such as “book chosen” in a reading experiment, might have a large number of possibilities. For other random factors, such as the clinic within which the experiment is conducted, the number of possibilities may be very limited, including only those in a narrow geographical region. The degree to which a researcher can generalize is dependent on the breadth of the universe of potential levels (e.g., range of book titles, number of different clinics). The authors should be careful not to interpret the absence of an interaction between the random and fixed factor as proving that the treatment works equally well in all levels of the random factor. Beyond the fact that a nonsignificant effect does not prove the null hypothesis to be true (see Desideratum 4), all generalizations should consider the universe of possible levels.

As with the randomized block design discussed in Desideratum 7, the choice of an appropriate error term depends on the assumptions made about the type of variables in the experiment. In mixed models, one of the factors is assumed to be a random variable and thus the error term for a fixed main effect is the interaction between the fixed and random variable. If there are few levels of the random factor, this error term will have few degrees of freedom and the associated *F* test will have little power.

In factorial experiments involving more than one random factor, there is generally no mean square that is a valid error term for the treatment effects of greatest interest. So-called *quasi-F* tests exist that can provide an approximate *F* test but, unless crucial to the question at hand, an experiment should involve no more than one random factor.

In some situations, the levels of the random variable might be viewed as replications of the fixed portion of the experiment in a different setting. This leads to separate analyses for each level of the random factor and the inability to attach a probability level to the possible presence of an interaction between the treatment and the random factor. When viewed as replications rather than parts of a single, large experiment with many levels of the random factor, the results may be summarized along with other experiments in a meta-analysis (see Chapter 19, this volume).

An ANOVA of a mixed design relies on relatively strict assumptions concerning the variability of subjects within levels and works best with balanced designs (i.e., an equal number of participants in each cell). In situations where there might be very different variances and systematically different sample sizes, authors should consider pursuing an analysis method that is better suited to handling the assumption violation (e.g., multilevel modeling; see Chapter 22, this volume).

9. Hierarchical Experiments

Hierarchical designs are similar to mixed models except that the random variable is typically nested within the levels of the fixed factor rather than crossed with them (see Desideratum 8). For a design that could be conducted either way, the mixed model is preferable in that the levels of the treatment variable would be repeated within each of the levels of the random variable. However, it is often the case that a level of the random variable can only be administered within one of the

treatments, or that the levels of the random variable are unique to a specific treatment. Consider a study to determine if listening to jazz, classical, or rock music differentially impacts studying. The random variable would be the particular musical pieces that are unique to a specific genre, where a control group with no music would probably be outside the hierarchical structure. Research involving subjects working together in teams are also typically hierarchical designs where the group is the random factor.

The authors should explain the rationale for selecting the nested variable, including the importance of being able to generalize the treatment effect averaged over all levels of the random variable. The breadth of the universe of levels from which the levels were chosen should be described including the procedures for selecting the specific levels.

Authors must acknowledge the dependence in the nested data by using an error term that includes this dependence. Main effects and interactions of fixed factors typically have error terms that involve the variability of the levels of the random factor rather than the variability of the subjects with the cells. In complex designs, special care should be taken by authors to clearly indicate the mean square that was used as the error term for each test along with its number of degrees of freedom.

The analysis of hierarchical designs with ANOVA is problematic under a number of commonly encountered situations. There is an assumption of homogeneity of residual variances within cells both within levels of the random and fixed factors that, when violated, can lead to inflation of Type I error rates. Equal sample size lessens the impact. If samples are systematically different in the number of subjects and the nature of the treatments leads one to believe the assumptions are likely to be violated, then the researchers should consider reanalysis using a multilevel modeling approach (see Chapter 22, this volume).

10. Incomplete and Other Complex Designs

Authors using Latin squares or other incomplete designs for the creation of treatment groups should provide a rationale to explain the advantage of the reduced number of treatment groups given the reduced amount of information provided. The rationale should include the authors' defense of no interactions between the factors that are incompletely crossed. In the Discussion section, the authors should acknowledge the confounding of the main effects with higher-order interactions inherent in the design.

Most importantly, the error term should be identified sufficiently to allow readers to understand the precise nature of the analysis conducted. For complex experimental designs, where a complete description may take extensive space, references should be provided.

11. Study Participants

The number of participants in each group or cell of the design should be clearly indicated. In addition, in complex experiments where the number of observations might be unclear, the authors should indicate the number of participants on which each mean, variance, etc. is based.

The authors must indicate the source of their participants in such a way that readers will be able to evaluate the generalizability of the findings. A statistically significant difference among the mean scores of different groups indicates that, within some specific population, the effects of the different treatments were unlikely to have all been equal. The concern of the reader is whether those differences apply in a setting of concern to them.

Among the characteristics that should be described are how participants were recruited, whether inducements were included to obtain participation, and any selection criteria that were used to

include or exclude participants. Sample specific demographic information should be included to understand better the implications of the study. Information about the number of individuals declining to participate and/or dropping out during the study should be provided so as to assess whether, for example, any differential attrition might bias treatment effect estimates.

The rationale for the number of participants used should be explained through a priori sample size determination (see also Chapter 27, this volume). Sample size should reflect the nature of the research question, the minimally important effect size, and the cost and availability of subjects. It is most important that the minimally important effect size be justified within the framework of the theory being investigated. Because of the direct relation between power and sample size, in the Discussion section the number of participants should be addressed vis-à-vis statistical significance and the strength of the observed relation.

12. Missing Data, Nonorthogonality, and Statistical Assumptions

Unless appropriate variables are available for imputing missing data, which are not the independent variables of the study, imputation is rarely appropriate in between-groups designs. Obviously, no imputation is appropriate for the level of the IV(s), and (unless an ANCOVA design is adopted) each individual contributes only a single continuous score so little is lost by excluding this case from the analysis. Keep in mind that a lack of independence between the main effects and interactions in factorial designs when cell sizes are disproportional, known as *nonorthogonality*, is often a function of missing data.

Whether nonorthogonality is a function of missing data or not, it impacts the analysis of the main effects (but not the interaction); thus, the strategy for addressing it should be indicated. Although Type I and Type III Sums of Squares are often reported because they are relatively straightforward and are the default in popular software packages, Type II Sums of Squares are usually the most appropriate. Type III Sums of Squares are most appropriate when an interaction is present, but when an interaction is present researchers should not be analyzing the main effects.

Parametric statistical procedures, such as ANOVA, have a number of assumptions on which their mathematical bases are built. In many cases assumptions are difficult, if not impossible, to completely satisfy, and the between-subjects ANOVA is generally not robust to violations of the assumptions. In ANOVA, the following assumptions are made: (1) individual errors are random and independent; (2) outcome variables are measured on at least an interval scale and are normally distributed within each level/cell of the design; and (3) variances within the levels/cells of the design are homogeneous.

The independence and randomness of errors is crucial for controlling the Type I error rate at the stated α level. Independence and randomness of errors are typically assumed to have been met during the design of the experiment, and this should be obvious from the researcher's description of how the participants were selected and assigned to groups. However, any potential independence issues must be dealt with appropriately in order to avoid any potential effects on the Type I error rate of the ANOVA F test. For example, if subjects are nested within a higher order factor (e.g., students within classrooms), then this nesting should be statistically accounted for (see Desideratum 8 of this chapter, and Chapter 22, this volume).

The assumption that the dependent variable be measured on an interval or ratio scale has attracted considerable debate over the years, with some maintaining that equal intervals are necessary in order to use ANOVA while others argue that the numbers themselves do not necessarily have any properties until one assumes those specific properties. We side with the latter, more liberal, position meaning that the analysis can be conducted with a scale of relatively unknown properties but

caution must be exercised in making the inferential leap to the dependent variable. Thus, if researchers have an ordered, multistep scale, they may proceed with ANOVA but should acknowledge that they are treating the numbers as if they were an interval scale. Further, the authors should clarify that the lack of fit affects the degree to which the research answers the questions that were raised.

The assumption that the scores on the DV are normally distributed within each population level or cell is routinely violated. Although simulation studies have demonstrated that violations seldom have any serious impact on the Type I error rate of the ANOVA F statistic, violations of normality can severely affect power and, if combined with violations of the variance homogeneity assumption, can drastically affect the Type I error rates and power of the between-groups ANOVA. Researchers should always evaluate and discuss the results of their assessment of normality, and adopt robust tests whenever assumptions are violated. If the normality assumption is violated, and the distributions have the same level of variability and are of the same shape, a nonparametric procedure (e.g., Kruskal–Wallis) is recommended. Significant increases in power can be observed when a nonparametric procedure is adopted, but care must be taken because for inferences to pertain to central tendency the variability and shapes of the distributions must be equal.

Homogeneity of population variances is an assumption that has minimal impact when violated, as long as equal sample sizes are used and the distributions are normal in shape. However, cell sizes are rarely equal and distributions are rarely normal, and thus researchers are expected to evaluate and report on the magnitude of the group or cell variances and adopt robust tests when necessary.

Several modified tests of significance can be used when heterogeneity of variance might be present, including tests by Brown and Forsythe, James, and Welch (factorial variants are also available). In some cases, differences in the variability of the outcome variable across the levels of the IV can also be an important research finding in and of itself, and can be verified using a test that evaluates the null hypothesis of equal population variances (e.g., Levene). Under simultaneous violation of the normality and variance homogeneity assumptions, heteroscedastic tests based on robust estimators (e.g., Welch test with trimmed means and Winsorized variances) or generalized linear models should be used (although caution is necessary when adopting generalized linear models because a specific relation is hypothesized between the central tendency and variability in most models).

It is important to highlight that adopting robust tests only if the data fail statistical tests of assumptions is generally not an effective strategy. In other words, the results of statistical tests of normality (e.g., Shapiro–Wilk) or variance homogeneity (e.g., Levene) are not good at predicting whether to use a traditional or robust procedure. The safest strategy is always to adopt robust tests; however, if this strategy is not adopted then robust tests should be used in any situation in which assumption violation is possible or expected.

13. Final ANOVA Model

After the design has been clarified and tests of assumptions are completed, researchers should indicate the nature of the omnibus ANOVA test being conducted (unless no omnibus test is being conducted; see Desideratum 14, in which case the nature of the specific hypothesis tests adopted is outlined). First, the acceptable probability of a Type I error (α) should be explicitly stated and justified. Just because .05 is used in most studies does not mean that it is appropriate for the researcher's specific hypotheses. For any test of significance reported, the authors should provide the test statistic (e.g., F), the source used as the error term, the degrees of freedom, and a statement about the probability level associated with the observed test statistic (in addition to the

strength of the effect, see Desideratum 15). ANOVA summary tables are space consuming and may not be needed for simple designs. More complex designs should have more complete reporting of the analysis so that the reader can understand a specific finding from within the context of the total experiment.

Each test of significance within an ANOVA depends on having a denominator (error term) of the F ratio whose expected value includes all of the terms in the expectation of the numerator except the one term whose existence is denied by the null hypothesis. While the correct error term is generally available in an experimental design text, there are some relatively common scenarios in which researchers use inappropriate error terms. The primary concern is related to the presence of random factors within an experiment. In designs with random factors, ignoring the proper error term in favor of the average within group variability (to capitalize on its greater number of degrees of freedom) can seriously inflate the Type I error rate reported.

The error term and degrees of freedom for the error term should always be reported. With the exception of sources of variation within incomplete designs, all error terms are either measures of variability between elements of some random factor (e.g., subjects or randomly selected levels) or the interaction of the elements of that random factor with a fixed factor. These sources are then sometimes pooled over random groups. The definition of an error term should include both the source or interacting sources and the levels over which it was pooled. The error term that is appropriate for the F test of the main effect or interaction should also be the basis of the standard error of the contrasts used in post hoc tests.

When robust tests are adopted, as recommended herein, it is important that researchers are very clear with regard to the nature of the analysis conducted. For example, Welch's heteroscedastic F test might be adopted if there was the potential for violating the variance homogeneity assumption, and references should be provided to support the use of the test within the given situation. If alternate estimators are used (e.g., trimmed means, Winsorized variances), then the details of the analysis (e.g., the proportion and nature of the trimming) should be explicit.

14. Follow-Up Strategies for Main Effects and Interactions

In one-way designs or for analyzing main effects in a factorial design, a multiple comparison procedure (MCP) can isolate exactly which levels of the IV differ significantly. The first step is for the researcher to define the error rate that is being controlled. In a factorial experiment, where several main effects and interactions might be tested, the authors should indicate if each main effect and interaction will be considered a family, with Type I error rate controlled separately at alpha (α) for each family, or if a more conservative level of control will be used.

Conducting each comparison with a Type I error rate of α (i.e., not controlling for multiplicity) implies control of the per-test error rate. This strategy is acceptable in exploratory studies, but can lead to an unacceptably high Type I error rate when several tests of significance are conducted. In most cases, it is recommended that researchers control the familywise error rate. However, if strict familywise error control is not required, but the researcher does not want to completely ignore the multiplicity issue, controlling the false discovery rate would be recommended for increasing power.

Follow-up tests can usually be adopted without having to conduct an omnibus F test; however, in more complex designs, such as a three way factorial designs, the omnibus tests provide a valuable tool for isolating statistically significant effects. Some multiple comparison procedures though, such as Fisher's least significant difference (LSD) procedure, require an omnibus test and therefore the preliminary test cannot be avoided in these situations. Of the commonly used MCP methods

(e.g., Tukey's honestly significant difference, Holm), all but the LSD and Newman–Keuls procedures provide adequate control of the familywise error rate without an omnibus test with more than three groups. Using an omnibus test as a gateway to conducting multiple comparisons will result in an unnecessarily conservative familywise Type I error rate.

The choice of major strategies to conduct MCPs should be based on the stated research hypotheses. The MCPs should have a one-to-one correspondence to the hypotheses. There are four general scenarios that commonly occur: (1) tests of planned, theory-derived contrasts; (2) tests of all possible pairs of means; (3) tests of each experimental group against a control group; and (4) tests that were not planned but rather suggested by the pattern found in the observed means.

The first three scenarios each have several MCP methods that may be used. The authors should clearly specify which method they chose for the analysis and why they chose this method. In general, Bonferroni's inequality (often called *Dunn's test*) is a potential method for controlling Type I error in any of the first three cases, however, methods with greater power exist for each strategy and therefore the Bonferroni approach should never be used. Holm's sequentially rejective procedure is a good option for any of the first three scenarios. It has greater power than the Bonferroni approach, maintains familywise error rates at α , can be used with any test statistic, and is available through most statistical software packages. Tukey's honestly significant difference (HSD) procedure is a popular procedure for analyzing all pairwise comparisons and Dunnett's procedure is a popular approach for analyzing all pairwise comparisons with a control. Both provide good familywise error control and relatively high power. As for the fourth scenario, involving comparisons/contrasts suggested by the observed data, Scheffé's test is required to adequately control Type I error. This is, however, the only scenario for which Scheffé's method is appropriate.

Interactions provide important information about the relation between a set of IVs and the DV, and thus should be analyzed in such a manner that the source of the significant interaction is unambiguous to the reader. Although these analyses are often time consuming, they are required in order to paint the reader an accurate picture of the nature of the interaction. It may be useful, as the first step in analyzing an interaction, to produce a visualization, using cell means in ANOVA or covariate-adjusted cell means in ANCOVA. Graphs, with the DV on the y -axis, and for a two-way design the predictors represented on the x -axis or as separate lines, often clearly indicate the source of the interaction, in which case statistical tests can be used to support these observations.

Post hoc analyses for interactions usually fall under one of two categories: (1) *simple effects*, or (2) *interaction contrasts*. Simple effect tests look at one effect at each level of another variable. For example, if a two-way interaction was statistically significant, a researcher might explore the effects of the first IV at each level of the second IV; or if a three-way interaction was statistically significant, a researcher might explore the interaction between the first two IVs at each level of the third IV (plus any required follow-ups to these analyses). Although this strategy is sometimes effective at understanding the nature of interactions, and is convenient for analyzing higher order (e.g., three-way) interactions, authors taking this approach forgo the opportunity to talk about how one effect differs across the levels of another variable. Interaction contrasts, on the other hand, which break down interactions into smaller contrasts, provide a much more direct interpretation of the nature of the interaction. For example, breaking a two-way interaction down to all 2×2 interactions will provide the reader with the source(s) of the statistically significant interaction. Multiplicity (e.g., familywise error rate) control is generally not necessary when researchers are conducting simple effect tests or interaction contrasts, unless inferential statements (i.e., hypotheses) regarding these tests are of interest to the researcher (rather than just trying to understand the nature of the interaction).

15. Effect Sizes and Confidence Intervals

A clear distinction should be made between the existence of a relation between the IV and the DV and the strength/direction of that relation. Tests of statistical significance, along with the probability levels, might lead one to infer that a non-chance relation exists, but this does not satisfy our responsibility to describe both the strength and the direction of the relation. Follow-up tests provide more information regarding the direction or nature of the relation, however the strength of the relationship remains to be established. Effect sizes are a necessary complement to the null hypothesis statistical tests that are used and also provide valuable information for researchers conducting meta-analyses (see Chapter 19, this volume).

An effect size can be any measure that relates to the hypothesis of interest (in this case, mean differences) and is not directly influenced by sample size. Therefore, raw mean differences, plots of the means, and so forth, are all valid and useful measures of the size of the effect. In fact, unstandardized measures like these are often preferable when the outcome variable is measured on a meaningful scale (although always be cautious that the variability of a DV can impact the interpretation of mean differences, etc.). A number of standardized measures of the size of the effect between the IV and DV have also been developed. For example, ω^2 (omega-squared) and η^2 (eta-squared) indicate the proportion of the variance present in the set of DV scores that is related to the differences in the levels of the IV (with ω^2 using unbiased variance components). Cohen's f is a standardized mean difference statistic that is also common. Because follow-up tests often involve comparing only two entities (e.g., two means or two mean differences), Cohen's standardized mean difference measure (d) is also popular. At present no one index is universally accepted as the standard.

Confidence intervals, when appropriate, provide important information regarding the parameters of interest (e.g., group mean differences). The width of the confidence interval should be interpreted relative to the nature of the outcome variable. Further, confidence intervals on effect sizes also add important information regarding the precision of the effect size.

16. Language Relative to Meaning and Generalizability

The logic of hypothesis testing presents some challenges in the use of language that authors should carefully consider. All statements should acknowledge the probabilistic nature of research in the social and behavioral sciences. In other words, we cannot prove anything; we infer there is a difference among populations because it is highly unlikely that the data we obtained would have happened if the null hypothesis were true. A retained null hypothesis should never be taken as proving the null hypothesis is true. When researchers would like to show that one method is as good as another, equivalence testing methods are required (see Desideratum 4).

In the literature that argues against the use of significance tests, one of the common complaints is the incorrect usage of language regarding the meaning of a *statistically significant* difference and a *not statistically significant* difference. The term *significant* should never be used in referring to the outcome of a test of significance without the additional word *statistical*. Statistical significance provides an unambiguous description of the results of a test of significance, whereas without the *statistical* qualifier we are left unsure whether the author is describing the outcome of the test of significance or arguing for the difference having real world importance.

Conclusions regarding the results should utilize both statistical significance and magnitude of the effects, and in turn generalizations from those results should be made with great caution. Statistically justified generalizations are made back to the theoretical population from which random sampling occurred. The problem is that almost all studies are completed with samples of convenience. We use

volunteers from our university or students in schools who have agreed to work with us. In the strictest sense, we have no meaningful population of subjects from which we sampled; we simply have available participants. All generalizing is then an extension beyond the actual data. Authors should make all generalizations as speculative. In the Methods section they should provide the reader with enough information about all of the conditions under which the experiment was conducted so that the reader is able to evaluate the similarity of the situation to one of concern to them.

References

- Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson Education.
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Monterey, CA: Brooks-Cole.
- Maxwell, S. E., & Delaney, H. D. (2003). *Designing experiments and analyzing data: A model comparison approach* (2nd ed.). Mahwah, NJ: Erlbaum.

2

Analysis of Variance *Repeated-Measures Designs*

Lisa M. Lix and H. J. Keselman

A repeated-measures design, also known as a *within-subjects design*, in which study participants are measured K times on the same dependent variable, is one of the most common research designs in the social, behavioral, and health sciences. The design occurs in both experimental and observational settings. Repeated measurements arise when a study participant is exposed to two or more experimental or treatment conditions (i.e., factor levels) such as different dosage levels of the same drug, or when a participant is observed at multiple points in time, leading to correlations among the outcome measurements. One advantage of this type of design is that, for a fixed sample size, it will generally result in greater precision of parameter estimates and more efficient inferential analyses than a between-subjects design. In addition, research questions about individual growth or maturation can only be effectively investigated in repeated-measures designs.

The repeated-measures analysis of variance (ANOVA) F test is the traditional procedure for testing hypotheses about within-subjects effects. This procedure makes stringent assumptions about the structure of the covariance matrix of the repeated measurements. Alternatives to the repeated-measures ANOVA F test may be more suitable for many of the data-analytic conditions encountered by researchers in the social, behavioral, and health sciences. Alternative procedures include: (a) an adjusted degrees of freedom (df) procedure, which modifies the df of the repeated-measures ANOVA critical value using information about the covariance matrix, (b) repeated-measures multivariate analysis of variance (MANOVA), which makes no assumptions about the structure of the covariance matrix of the repeated measurements (except across any between-subjects factors in the design), (c) the multiple regression model, which allows the researcher to characterize the variances and covariances of the repeated measurements using a small number of parameters, (d) the random-effects (e.g., mixed-effects) model, which allows the researcher to describe and test subject-specific variation in repeated-measures data, (e) the latent growth curve model, which is based on structural equation modeling (SEM) and discussed in Chapter 13, and (f) approximate df procedures, which do not assume that the data follow a normal distribution or that covariances are equal (i.e., homogeneous) across any between-subjects factors in the design.

Several comprehensive resources that describe procedures for the analysis of repeated-measures data are available. These include Fitzmaurice, Davidian, Verbeke, and Molenberghs (2009), Hedeker and Gibbons (2006), Hoffman (2015), and Singer and Willett (2003).

Desiderata for the analysis of repeated-measures data are given in Table 2.1. A detailed discussion of each item is provided below.

Table 2.1 Desiderata for the Analysis of Repeated Measurements.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. The characteristics of the repeated-measures design are specified (i.e., number of within- and between-subjects factors, number of levels of each, spacing between the measurement occasions).	I, M
2. Issues of statistical power have been considered and the sample size is reported.	M
3. The number and type of dependent (i.e., response) variable(s) is specified. Covariates are considered for inclusion in the analysis.	M
4. Assumptions about the distribution of the dependent variable(s) are evaluated and an appropriate test procedure is selected.	M
5. Assumptions about the covariance structure of the repeated measurements are evaluated and used to guide the selection of a test procedure.	M
6. The pattern and rate of missing observations is considered. The method adopted to handle missing observations is identified.	M,R
7. The method used to conduct <i>a priori</i> or post hoc multiple comparisons of the repeated measurements is specified. The method adopted for testing multiple dependent variables, if present, is specified.	M
8. The name and version of the selected software package is reported.	M, R
9. Exploratory analyses of the repeated-measures data are summarized.	R
10. The results for omnibus tests and multiple comparisons on the repeated measurements are reported. The criterion used to assess statistical significance is specified. Evaluations of model fit are described.	R
11. Consideration is given to reporting effect sizes and confidence intervals.	R
12. The strengths and limitations of a repeated-measures design and the selected method of analysis are considered and threats to the validity of study findings are discussed.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

1. Types of Repeated-Measures Designs

Information about the characteristics of a repeated-measures design is used to assess the overall appropriateness and validity of the hypothesis-testing strategy. The simplest repeated-measures design is one in which a single group of study participants is measured on one dependent variable at two or more occasions or for two or more experimental or treatment conditions. Consider an example in which a study cohort is observed repeatedly after being introduced to a new therapeutic treatment. Suppose the researcher is interested in investigating the treatment's effect on the quality of life of study participants. An appropriate null hypothesis for such a design is that there is no change in average quality of life ratings over time. If this omnibus hypothesis is rejected, multiple comparisons might be conducted to test for a mean difference between pairs of measurement occasions. *A priori* contrasts among the measurement occasions could be conducted instead of a test of the omnibus hypothesis (see Desideratum 7).

Factorial repeated-measures designs contain two or more repeated-measures factors. The simplest factorial design is one in which a single group of study participants is measured on a single dependent variable for each possible combination of the levels of two factors. For example, suppose that a researcher investigates psychological well-being of a cohort exposed to experimental stimuli that represent all combinations of sex (male, female) and facial expression (positive, neutral, negative). The dependent variable in this example is psychological well-being. The

null hypothesis for the within-subjects interaction effect is that the effect of sex of the stimuli on psychological well-being is constant at each level of the facial expression factor. If the null hypothesis is rejected, multiple interaction contrasts might be conducted to identify the combination of factor levels that contribute to rejection of the omnibus hypothesis. If the interaction effect is not significant, the researcher can choose to test the main effects (i.e., sex, facial expression; see Desideratum 10).

A mixed design, also referred to as a *split-plot* repeated-measures design, contains both between-subjects and within-subjects factors. The simplest mixed design contains a single within-subjects factor, a single between-subjects factor, and a single outcome variable. For example, study participants might be randomly assigned to control and intervention groups prior to being measured at successive points in time on their reading comprehension. In a mixed design, the researcher is primarily interested in testing whether there is a group-by-time interaction, that is, whether the change over time on the dependent variable (e.g., reading comprehension) is the same for control and intervention groups, although main effects will be of interest if the interaction is not significant (see Desideratum 10).

2. Sample Size and Statistical Power

Statistical power, the probability that an effect will be detected when it exists in the population, and sample size (N) must be considered early on in the design of a study. While the conventional ANOVA procedure will often have the greatest power to detect a within-subjects effect (Fang, Brooks, Rizzo, Epsy, & Barcikowski, 2008), it is unlikely to be the optimal choice because it rests on such a stringent set of assumptions.

Calculating the sample size to achieve a desired level of statistical power requires information about the pattern and magnitude of the within- and between-subjects effect(s), the variances and covariances at each measurement occasion, the number of measurement occasions, the level of significance (α), and the choice of analysis procedures (Guo, Logan, Glueck, & Muller, 2013). Information about the magnitude of effects, as well as the pattern of variances and covariances of the repeated measurements, can often be obtained from previous research or from a pilot study. However the use of existing data to estimate the covariance matrix and/or effect size has the consequence that power becomes a random variable; methods have therefore been developed to establish accurate confidence intervals for power in such situations (Gribbin, Chi, Stewart, & Muller, 2013), to guide researchers in selecting the correct sample size for a study.

Many statistical software packages can be used to calculate the sample size requirements for repeated-measures designs. For complex designs that have multiple dependent variables, multiple independent variables, and/or clustering effects, the estimation of sample size requirements may be less straightforward. Computer simulation may be useful to generate estimates of power and the required sample (Arnold, Hogan, Colford, & Hubbard, 2011). The researcher is advised to consult with a statistician for complex designs.

However, in some cases the available sample size may dictate the choice of analysis procedures. For example, if N is less than the number of measurement occasions (K), the repeated-measures MANOVA procedure cannot be used to test within-subjects effects. Furthermore, if the ratio N/K is small, the covariance parameter estimates may be unstable; the latent growth curve model would not be a good choice in such a situation. In mixed designs, the ratio of the group sizes, in addition to the total sample size, is a consideration in the choice of procedures; if group sizes are unequal and equality (i.e., homogeneity) of the group covariances is not a tenable assumption, then the repeated-measures ANOVA and MANOVA procedures may result in invalid inferences (e.g., too many false rejections of null hypotheses).

3. Number and Type of Dependent Variables

Repeated-measures designs may be either univariate or multivariate in nature. A multivariate repeated-measures design is one in which measurements are obtained from study participants on P dependent variables at each occasion. In multivariate data there are two sources of correlation: (a) within-individual within-variable correlation, and (b) within-individual between-variable correlation. The latter arises because the measurements obtained on the dependent variables at a single occasion are almost always related.

One approach to analyze multivariate repeated-measures data is to conduct P tests of within-subjects effects, one for each dependent variable. This method can be substantially less powerful than a multivariate analysis, which simultaneously tests within-subjects effects for the set of P outcomes. Several procedures have been proposed to test multivariate within-subjects main and interaction effects (Lix & Lloyd, 2007; Vallejo, Fidalgo, & Fernandez, 2001; Verbeke, Fieuws, Molenberghs, & Davidian, 2014). Two conventional procedures are the *doubly multivariate model* (DMM) and *multivariate mixed model* (MMM) procedures, which are extensions of repeated-measures MANOVA and ANOVA, respectively, to the case of two or more dependent variables. The choice between these two procedures is a function of sample size and one's assumptions about the data. The DMM cannot be applied to datasets in which $N/(P \times K)$ is less than one. Moreover, when this ratio is small, covariance parameter estimates may be unstable. The MMM procedure makes stringent assumptions about the covariance structure of the repeated measurements and dependent variables. When these assumptions are not satisfied, the MMM will result in invalid inferences.

The multiple regression model (MRM) procedure is an appealing alternative to these conventional procedures. It allows the researcher to define the covariance matrix of the repeated measurements and dependent variables using a small number of parameters. A parsimonious (i.e., simple) structure for the MRM is a separable structure, in which the covariance matrix of the repeated measurements is assumed to be the same for each dependent variable. It is advantageous to assume a separable covariance structure when sample size is small because it requires estimation of fewer covariance parameters than when an unstructured covariance is assumed and therefore the estimates will be more stable.

In either univariate or multivariate repeated-measures data, the dependent variables may have a continuous or discrete scale. For the latter, the outcome might be the presence (or absence) of a response or a count of the number of times a response occurs. Generalized linear models are a unified class of models for the analysis of discrete data. They have been extended to the case of correlated observations. Binary repeated measurements can be analyzed using an extension of logistic regression, repeated counts of rare events can be analyzed using an extension of Poisson regression, and repeated ordinal measurements can be analyzed using an extension of multinomial regression for repeated-measures data (Hedeker & Gibbons, 2006). Two different types of generalized linear models for repeated measurements are marginal models and random-effects models. The choice between these two approaches is largely a function of the research purpose; a marginal model is used when the researcher is interested in making inferences about the average response in the population while the random-effects model is used when the researcher is interested in making inferences about the response of the average individual in the population.

4. Distributional Assumptions

Repeated-measures ANOVA, MANOVA, and multiple regression model (MRM) procedures rest on the assumption of multivariate normality. If the repeated measurements are distributed as multivariate normal, then the data for each measurement occasion are normally distributed and the joint distribution of the data for all measurement occasions are normally distributed. However, even

when the data for each measurement occasion are normally distributed, the set of measurements might not follow a multivariate normal distribution (Keselman, 2005; Looney, 1995).

Assessing potential departures from a multivariate normal distribution is critical to selecting a valid analysis procedure. The researcher can compute measures of skewness (symmetry) and kurtosis (tail weight) for the marginal distributions (i.e., for each measurement occasion), as well as measures of multivariate skewness and kurtosis. Values near zero (assuming an adjusted measure of kurtosis) are indicative of a normal distribution. Many tests of univariate and multivariate normality have been proposed (Zhou & Shao, 2014). However, these tests may be sensitive to even slight departures from a normal distribution, and therefore may not be useful for decision making. Data exploration tools, such as normal probability plots, might be more useful for assessing departures from a multivariate normal distribution; these are discussed in further detail in Desideratum 9. Alternatively, one may simply bypass these assessments of the data distribution in favor of a test procedure that is robust (i.e., insensitive) to departures from multivariate normality.

Procedures such as the repeated-measures ANOVA and MANOVA are sensitive to the presence of outliers in the data distribution (Oberfeld & Franke, 2013). Outliers inflate the standard error of the mean resulting in reduced sensitivity to detect within-subjects effects. One approach to overcome the biasing effects of non-normality is to adopt robust measures of central tendency and variability. Trimmed means and Winsorized variances have been extensively studied as alternatives to conventional least-squares means and variances in the analysis of repeated-measures data (Keselman, Wilcox, & Lix, 2003). A trimmed mean is obtained by removing an *a priori* determined percentage of the largest and smallest observations at each measurement occasion and computing the mean of the remaining observations. A commonly-recommended trimming percentage is 20% in each tail of the distribution. The trimmed mean will have a smaller standard error than the least-squares mean when the data are sampled from a heavy-tailed distribution (i.e., a distribution containing outliers or extreme scores). In Winsorizing, the smallest non-trimmed score replaces the scores trimmed from the lower tail of the distribution and the largest non-trimmed score replaces the scores removed from the upper tail. These non-trimmed and replaced scores together are called Winsorized scores. A Winsorized mean is calculated by applying the usual formula for the mean to the Winsorized scores; a Winsorized variance is calculated using the sum of squared deviations of Winsorized scores from the Winsorized mean. The Winsorized variance is used instead of the trimmed variance, because the standard error of a trimmed mean is a function of the Winsorized variance. Test procedures based on robust estimators have demonstrated good performance (i.e., accurate Type I rates and acceptable levels of statistical power to detect non-null treatment effects) for analyzing repeated-measures data. Computer programs that use the trimmed mean and Winsorized variance to test within-subjects effects are discussed in Desideratum 8.

Another approach to deal with the biasing effects of non-normality is to transform the data prior to analyzing it. Rank transform procedures, in which observations are ranked prior to applying an existing procedure for analyzing repeated measurements, are appealing because they can be easily implemented using existing statistical software packages (Conover & Iman, 1981). One limitation is that they cannot be applied to tests of within-case interaction effects, because the ranks are not a linear function of the original observations. Therefore, ranking may introduce additional effects into the statistical model that were not present in the original data. Ranking may also alter the pattern of correlations among the repeated measurements, which can be particularly problematic for the repeated-measures ANOVA procedure because it makes specific assumptions about the correlation structure of the data. Thus, rank transform procedures, while insensitive to departures from a normal distribution of responses, must be used with caution.

Other normalizing transformations may be applied to the data, such as the logarithmic transformation. However, the primary problem with applying a transformation to produce

a more normal distribution is that inferences must now be made on the means of the transformed data. Once the analysis has been conducted, back-transforming to the data into the original scale of measurement is not possible. A common normalizing transformation is the natural logarithm.

The generalized linear model, either with or without random effects, provides a comprehensive framework for modeling discrete and continuous data that do not follow a normal distribution. Common choices include the Poisson distribution for count data, Bernoulli distribution for binary data, and gamma distribution for skewed continuous data; all of these distributions are encompassed within the exponential family of distributions. Other families of distributions may be adopted if these conventional choices do not provide a good fit to the data (Arnau, Bono, Blanca, & Bendayan, 2012; Molenberghs, Verbeke, Demetrio, & Vieira, 2010), as judged by goodness of fit statistics such as Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC). The BIC penalizes models more severely for the number of parameter estimates than does the AIC, therefore the latter is usually favored over the former.

A nonparametric bootstrap resampling method has also been proposed for the analysis of non-normal repeated-measures data (Berkovits, Hancock, & Nevitt, 2000). Under this methodology, a normal-theory procedure is used to test within-subjects effects; however, the critical value for evaluating statistical significance is based on the empirical sampling distribution of the test statistic rather than a theoretical critical value (e.g., a critical value from an F distribution).

5. Covariance Structure of the Repeated Measurements

Procedures for analyzing repeated-measures data vary widely in their assumptions about the structure of the covariance matrix; evaluation of the covariance structure is therefore critical to the selection of a valid method of analysis. The repeated-measures ANOVA procedure assumes that the covariance matrix of the repeated measurements has a spherical structure. For *sphericity* to be satisfied, the population variances of the differences between pairs of repeated-measures factor levels must be equal. Furthermore, for mixed designs, the more stringent assumption of *multisample sphericity* must be satisfied; this requires equality of the common variance of the pairwise repeated-measures differences across levels of the between-subjects factor. The sphericity assumption is not likely to be satisfied in data arising in social, behavioral, and health sciences research. For example, when measurements are obtained at multiple points in time, it is often the case that the variance increases over time. Moreover, a test of the sphericity assumption is sensitive to departures from a multivariate normal distribution. Therefore, the repeated-measures ANOVA procedure is not routinely recommended in practice.

The *approximate df* ANOVA procedure is one alternative when sphericity is not a tenable assumption (Brunner, Bathke, & Placzek, 2012; Skene & Kenward, 2010a). The *repeated measure* MANOVA procedure is another alternative; it does not make any assumptions about the structure of the common covariance matrix of the repeated measurements. However, both the adjusted df ANOVA and repeated-measures MANOVA procedures do assume homogeneity of the group covariance matrices across any between-subjects factor levels in the design (Brunner et al., 2012). These procedures are not robust to departures from covariance homogeneity, particularly when group sizes are unequal. If the group with the smallest sample size exhibits a larger degree of variability among the covariances than the group with the largest sample size, then tests of within-subjects effects will tend to produce liberal Type I error rates, above the nominal α level (i.e., too often inferring there are within-subjects effects when none are present). Conversely, if the group with the smallest sample size exhibits the smallest degree of variability of the covariances, then tests of within-subjects effects will tend to produce conservative error rates, below the nominal α level (i.e., too often failing to detect real population effects). Unfortunately, a likelihood ratio procedure

to test the null hypothesis of covariance homogeneity is sensitive to departures from a multivariate normal distribution, as well as to small sample sizes.

When it is not reasonable to assume that covariances are homogeneous, the researcher is recommended to bypass these analysis procedures in favor of an approximate df procedure (Keselman, Algina, Lix, Wilcox, & Deering, 2008). The approximate df procedure, which is a multivariate and multi-group generalization of the non-pooled two-group t test, has been extensively studied for both univariate and multivariate repeated-measures designs when covariances are heterogeneous. It will result in valid inferences about within-subjects effects provided that sample size is not too small. However, the approximate df procedure does assume that the repeated-measures data follows a multivariate normal distribution. When multivariate normality is not a tenable assumption, then the approximate df procedure should be implemented by substituting trimmed means and Winsorized variances for the usual least-squares means and variances. A computer program for this procedure is described in Desideratum 8.

The multiple regression model (MRM) procedure allows the researcher to model the covariance matrix of the repeated measurements in terms of a small number of parameters (Skene & Kenward, 2010b). Heterogeneous covariance structures can also be accommodated for mixed designs if homogeneity of group covariances is not a tenable assumption. There are several different covariance structures that can be fit to one's data. Autoregressive and Toeplitz structures assume that the correlation among repeated measurements is a function of the lag, or interval, between two measurement occasions. Some covariance structures assume that the variances of the measurement occasions are constant, while other structures allow for heterogeneous variances. For example, the random coefficients structure is a flexible structure that models subject-specific variation characterized by non-constant variances and non-constant correlations.

When a parsimonious covariance structure is specified for the repeated measurements, the MRM procedure will result in a more powerful test of within-subjects effects than the repeated-measures MANOVA procedure. However, if the covariance structure is incorrectly specified, tests of within-subjects effects may be biased, resulting in erroneous inferences.

The latent growth curve model also allows the researcher to explicitly model the correlation structure of the repeated measurements. Moreover, the SEM approach allows the researcher to separate the model for the mean structure and the covariance structure of the data, which can be advantageous for describing the characteristics of one's data (Curran, Obeidat, & Losardo, 2010).

Graphic techniques and summary statistics to aid in selecting the initial model(s) for the covariance structure are described in Desideratum 9. Measures of model fit and/or inferential analyses are also critical in order to correctly specify the covariance structure (Liu, Rovine, & Molenaar, 2012). If the candidate covariance structures are nested, then a likelihood ratio test can be used to select one of these structures as the final model. Two covariance structures are nested if one is a special case of another. For example, a *compound symmetric* covariance structure, which assumes that all variances are equal and all covariances are equal, is a special case of an unstructured covariance model, which does not assume that either variances or covariances are equal. Caution is advised when adopting the likelihood ratio test because it is sensitive to multivariate non-normality and small sample size. The AIC and BIC can be used to compare model fit for non-nested covariance structures. These information criteria should only be used for comparing the covariance structures of models that contain the same regression parameters, so that a direct assessment of the effect of the covariance structure on the fit is obtained.

6. Missing Observations and Loss to Follow-Up

Missing data are a concern in repeated-measures designs because of the potential loss of statistical efficiency and/or bias in parameter estimates due to differences between observed and unobserved data (Vallejo, Fernandez, Livacic-Rojas, & Tuero-Herrero, 2011). Although researchers may devote

substantial effort to reduce the amount of missing data, some loss of data is inevitable. Therefore, an assessment of the amount and type of missing data is essential in a study involving repeated measurements.

Missing data can be either monotone or intermittent. Both patterns can appear in the same dataset. A *monotone*, or drop-out, pattern arises if a participant is observed on a particular occasion but not on subsequent occasions. Study drop-out may arise for a number of reasons, including death or illness, or lack of interest in continuing a study. An *intermittent* pattern is one in which there are “holes” in the data because a study participant will have at least one observed value following a missing observation. Overall, as the rate of missingness (i.e., the proportion of missing observations to the total number of observations in the dataset) increases, statistical efficiency decreases and the probability of biased inference increases.

Repeated-measures ANOVA, adjusted df ANOVA, and repeated-measures MANOVA procedures assume a complete set of measurements for each study participant. If data are incomplete, the researcher is faced with the following choices. First, one could simply exclude from the analysis all study participants with at least one missing observation (so called *casewise* or *listwise* deletion). The final sample size available for analysis can be extremely small if the rate of missing observations is large. A second option is to choose an analysis procedure based on maximum likelihood estimation (e.g., *full information* maximum likelihood estimation), which does not result in deletion of participants with missing observations. Another choice is to impute missing values, using single or multiple imputation methods, in order to obtain a complete data set for subsequent analysis.

Examples of single imputation methods include mean substitution or last observation carried forward. These methods are not widely recommended, particularly when the rate of missing observations is large, because they do not account for random variation in the missing observations. Single imputation will therefore result in model parameters with underestimated error variances. Multiple imputation, the preferred approach (Little & Rubin, 2002), generates M plausible sets of values for the missing observations, yielding M pseudo-complete datasets that are analyzed using complete-case methods and whose results are combined using simple arithmetic formulae. There is no single value of M that is recommended in practice, although Schafer (1999) suggested that between three and ten imputations will likely be sufficient for the majority of missing data problems. The value of M depends on the rate of missing observations. One strategy for choosing M is to conduct several sets of M imputations, starting with a small value of M and evaluating whether parameter estimates are relatively stable across these independent sets of imputations. If the estimates demonstrate wide variability, then M should be increased and the stability of the estimates re-evaluated.

Imputation-based analyses will result in unbiased tests of within-subjects effects only if the missing data are ignorable (Vallejo et al., 2011). There are three mechanisms by which data may be incomplete (Little & Rubin, 2002): (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR). MCAR means the probability that an observation is missing is independent of either observed or unobserved responses. MAR means the probability that an observation is missing depends only on the pattern of observed responses. All other missing data mechanisms are MNAR, or non-ignorable. Unfortunately, there are no formal tests of the null hypothesis that the missing data follow a MAR pattern instead of a MNAR pattern. Pattern selection models or pattern mixture models are recommended to reduce biases when the data are assumed to be MNAR (Hedeker & Gibbons, 1997). Pattern selection models rely on multiple imputations, under a variety of assumptions about the missing data mechanism, to test effects and/or estimate model parameters. Pattern mixture models develop a categorical predictor variable for the different patterns of missing data and this predictor variable is included in the statistical model for testing within-subjects effects. No single missing data model can be uniformly recommended to reduce bias in parameter estimates; the choice depends on the type and rate of

missing observations, the number of measurement occasions, and the magnitude of the within-subjects effects. Shen, Beunckens, Mallinckrodt, and Molenberghs (2006) proposed using sensitivity analyses when the missingness is assumed to be non-ignorable, to assess whether the findings for different missing data models produce consistent results. Sensitivity analysis techniques can also be used to identify influential observations in analyses of missing data.

7. Multiple-Comparison Procedures and Multiple Testing Strategies

When conducting post hoc or *a priori* multiple comparisons, such as pairwise contrasts among within-subjects factor levels, the researcher will typically wish adopt a procedure to control the familywise error rate (FWER), the probability of committing at least one Type I error, for the set of tests. The well-known Dunn–Bonferroni procedure conducts each of C comparisons at the α/C level of significance. The Bonferroni method is simple to implement, but may not be as powerful as other procedures, such as Hochberg's (1988) step-up procedure. This procedure orders the p -values from smallest to largest so that $p_{(1)} \leq p_{(2)} \dots \leq p_{(C)}$, to test the corresponding hypotheses $H_{(1)}, \dots, H_{(C)}$. The sequence of testing begins with the largest p -value, $p_{(C)}$, which is compared to α . Once a hypothesis is rejected, then all hypotheses with smaller p -values are also rejected by implication. For example, if $p_{(C)} \leq \alpha$, then all C hypotheses are rejected. If the null hypothesis corresponding to the largest p -value, $H_{(C)}$, is accepted, the next p -value, $p_{(C-1)}$, is evaluated using the $\alpha/2$ criterion. More generally, the decision rule is to reject $H_{(w')}$ ($w' \leq w$; $w = C, \dots, 1$) if $p_{(w)} \leq \alpha/(C - w + 1)$. An assumption underlying Hochberg's procedure is that the tests are independent, which is unlikely to be satisfied in a repeated-measures design. However, Hochberg's procedure will control the FWER for several situations of dependent tests, making this procedure applicable to most multiple-comparison situations that social scientists might encounter (Sarkar & Chang, 1997). Multiple-comparison procedures for correlated data are discussed later in this Desideratum.

In factorial and mixed designs, multiple comparisons to probe interaction effects should be conducted using interaction contrasts (Lix & Keselman, 1996). Tests of simple main effects, which involve examining the effects of one factor at a particular level of the second factor, may also assist researchers in examining the interaction. A significant interaction implies that at least one contrast among the levels of one factor is different at two or more levels of the second factor. Tetrad contrasts are one type of interaction contrasts that are a direct extension of pairwise contrasts for probing marginal (i.e., main) effects. In a two-way design, a tetrad contrast involves testing for the presence of an interaction between rows and columns in a 2×2 sub-matrix of the data matrix, or in other words, of testing for a difference between two pairwise differences. Control of the FWER for a set of tetrad contrasts can be achieved using an appropriate multiple-comparison procedure as described previously.

In multivariate repeated-measures data, one strategy to conduct multiple comparisons is to follow a significant omnibus multivariate effect with post hoc multivariate multiple comparisons. For example, a significant multivariate within-subjects interaction indicates that the profiles of the repeated measurements are not parallel for two or more levels of the between-subjects factor for some linear combination of the outcome variables. Multivariate interaction contrasts are an appropriate choice for probing this effect.

There are many kinds of comparisons that might be tested in a multivariate design. Bird and Hadzi-Pavlovic (1983) distinguished among *strongly restricted contrasts*, which are defined for between- and/or within-subjects factor levels on a single outcome variable, and *moderately restricted contrasts*, which are defined for between-subjects and/or within-subjects factor levels for two or more outcome variables. A third type, the *unrestricted contrast*, is defined as the maximum contrast for the first linear discriminant function, that is, the linear combination of coefficients that

maximizes the distance between the means of the outcome variables. Unrestricted contrasts can be difficult to interpret because the coefficients are usually fractional, while strongly restricted contrasts are the easiest to interpret because they focus on only a single outcome variable. At the same time, a simultaneous test procedure (e.g., Bonferroni) to control the FWER for all possible strongly restricted contrasts will have very low power to detect significant effects because it uses a stringent criterion to evaluate each test statistic. A more powerful approach is to conduct a small set of *a priori* multivariate contrasts on the between-subjects or within-subjects factor levels for the set of dependent variables, using a stepwise multiple-comparison procedure, such as Hochberg's (1988) procedure, to control the FWER for the set of tests.

Another approach to probe multivariate repeated-measures data is to conduct tests of within-subjects effects for each of the P outcomes, adopting a significance criterion to control the FWER that is also adjusted for the correlation among the outcomes. The Bonferroni method and its stepwise counterparts assume that the outcomes are independent and will therefore result in conservative tests of within-subjects effects on the P outcomes, particularly when P is large. Alternate approaches that adjust for correlation include Roy's (1958) step-down analysis and resampling-based methods.

In a step-down analysis, the researcher rank orders the outcome variables in descending order of importance and then conducts tests of within-subjects effects using an analysis of covariance (ANCOVA) approach in which higher-ranked outcome variables serve as covariates for tests on lower-ranked variables. Under the null hypothesis and assuming that the data are normally distributed, the step-down test statistics, F_l ($l = 1, \dots, P$) and p -values, p_l , are conditionally independent. The FWER for the set of step-down tests is controlled to α using a multiple-comparison procedure such as Hochberg's (1988) method. A step-down analysis is an appropriate method if the researcher is able to specify an *a priori* ordering of the outcome variables; this is often the case when some outcomes have a greater theoretical importance to the researcher than others.

Westfall and Young (1993) described a step-down resampling-based multiple testing procedure that also adjusts for the correlation among multiple outcomes. Their procedure uses a permutation method, in which the observations are reshuffled or re-randomized. The permutation procedure is implemented as follows: A permuted dataset is obtained by reshuffling the original observations. A test of the within-subjects effect is computed for each of the P outcome variables. The test statistic in the permuted dataset that corresponds to the maximum test statistic in the original dataset is used to evaluate statistical significance for each of the P outcome variables. This process is repeated B times. The p -value for the m th outcome variable ($m = 1, \dots, P$) is the proportion of permutations in which the maximal criterion exceeds the value of the m th test statistic in the original dataset. A critical issue in implementing this multiple testing procedure is ensuring that the data are re-randomized correctly. For example, to test the within-subjects interaction effects in a mixed design the data must be doubly randomized, that is, reshuffled among rows as well as among columns of the original data matrix.

8. Software Choices

Procedures for the analysis of repeated-measures data are available in software packages commonly used by researchers in the social, behavioral, and health sciences including SPSS, SAS, Stata, and R. Reporting the name and version of statistical software is recommended because not all packages will rely on the same default options for estimating model parameters or testing within-subjects effects. Options available for imputing missing values and conducting computationally-intensive re-sampling techniques may not be the same in all software packages.

Syntax to implement repeated-measures ANOVA, MANOVA, and multiple regression model (MRM) procedures are described in a number of sources. Singer and Willett (2003) and Newsom,

Jones, and Hofer (2012) are just two examples of the resources that offer downloadable programs to analyze within-subjects effects in multiple software packages.

The approximate df procedure for testing within-subjects effects in the presence of covariance heterogeneity is not currently available in commercial statistical software packages. A program written in the SAS/IML language to implement this solution is available by contacting the first author. Numeric examples that demonstrate this software for a variety of research designs are available (Keselman et al., 2008), along with documentation about its implementation. Tests can be conducted using least-squares means and variances or trimmed means and Winsorized variances. The program will evaluate statistical significance of tests of within-subjects effects using either a critical value from an F distribution or a bootstrap critical value. As well, it will compute robust effect size estimates and robust confidence intervals; these are described in more detail in Desideratum 11.

9. Exploratory Analysis Techniques

Graphic techniques and summary statistics are used to evaluate the tenability of derivational assumptions that underlie different methods for the analysis of repeated-measures data and to aid in the selection of an appropriate model for the covariance structure of the repeated measurements when appropriate. The results of exploratory analyses should be summarized, because they provide an assurance that the choice of analysis procedures is justified.

Profile plots of the data for individual study participants are used to assess the magnitude of subject-specific variation in the data and whether that variation is increasing or decreasing across measurement occasions, which could result in violations of the assumption of sphericity. Scatter plots for pairs of measurement occasions can aid in the identification of potential outliers or influential observations. A *normal probability plot*, or *normal quantile plot*, is a scatter plot of the percentiles of the data versus the percentiles of a population from a normal distribution. If the data do come from a normally-distributed population, the resulting points should fall closely along a straight line.

Summary statistics such as correlation coefficients and variances can aid in the selection of a model for the covariance structure of the repeated measurements. The *correlogram*, which plots the average correlation among the measurement occasions against the number of lags (h) between the occasions ($h = 1, \dots, K - 1$), can also be used for this purpose. Finally, change scores between pairs of measurement occasions might be useful for identifying informative post hoc contrasts for probing within-subjects effects.

10. Reporting Test Statistic Results

For completeness, the test procedure(s) used to conduct all analyses should be specified. For example, when the repeated-measures MANOVA procedure is used to analyze data arising from a multi-group mixed design, there are four different test statistics that can be used to test the within-subjects interaction: the Pillai–Bartlett trace, Roy's largest root criterion, Wilks's lambda, and the Hotelling–Lawley trace. These statistics represent different ways of summarizing multivariate data. When the design contains two groups, all of these tests reduce to Hotelling's T^2 , a multivariate extension of the two-sample t statistic. All four multivariate criteria rest on the assumption of a normal distribution of responses and homogeneity of group covariances. Olson (1976) found the Pillai–Bartlett trace to be the most robust of the four tests when the multivariate normality assumption is not tenable, and is sometimes preferred for this reason.

For the MRM, several statistics are available to test hypotheses about covariance structures and within-subjects effects. Tests about covariance parameters can be made using a Wald z statistic,

which is constructed as the parameter estimate divided by its asymptotic standard error. However, this test statistic can produce erroneous results when sample size is small. The likelihood ratio test for comparing nested covariance structures, which asymptotically follows a χ^2 distribution when the data are normally distributed, is sensitive to small sample sizes. Specifically, Type I error rates may exceed the nominal α . For testing hypotheses about within-subjects main and interaction effects, a Wald F statistic can be used; it has good performance properties in large samples (Gomez, Schaafje, & Fellingham, 2005). Improved performance in small sample sizes can be obtained either by adjusting the df of the test statistic or modifying the test statistic value (Skene & Kenward, 2010a, b); this option is available in SAS software.

11. Effect Sizes and Confidence Intervals

An effect size describes the magnitude of a treatment effect (see Chapter 6). Reporting effect sizes in addition to hypothesis testing results is required in some journal editorial policies and is supported by the American Psychological Association's Task Force on Statistical Inference. One commonly-reported measure of effect size is Cohen's d . In a repeated-measures design, this measure is computed as the standardized difference of the means for two within-subjects factor levels, taking into account the correlation between the measurement occasions. A confidence interval should also be reported for an effect size to provide information about the precision of the estimate. The noncentral t distribution is used to construct a confidence interval when the data are normally distributed. Effect size measures need not be limited to the case of only two within-subjects factor levels; Keselman et al. (2008) discussed this issue in detail.

When the data are not normally distributed, the coverage probability of the confidence interval is poor (Algina, Keselman, & Penfield, 2005), and may become worse as the correlation among the measurement occasions increases. One option is to use an empirical method, such as the bootstrap, to construct a confidence interval. A bootstrap dataset is obtained by randomly sampling with replacement from the original data. An effect size measure is computed from the bootstrapped dataset. This process is repeated B times. The B effect sizes are ranked in ascending order. The $B \times (\alpha/2)$ and $B \times (1 - \alpha/2)$ observations of the empirical distribution represent the upper and lower limits of the $100 \times (1 - \alpha)\%$ confidence interval, respectively. A measure of effect size that is insensitive to departures from a multivariate normal distribution can be obtained by using the trimmed mean and Winsorized variance in place of the usual least-squares mean and variance.

Cohen's effect size assumes homogeneity of group covariances in mixed designs, because the denominator, or "standardizer," of the effect size is based on an estimate of error variance that averages across levels of the between-subjects factor. As a result, when covariances are heterogeneous and group sizes are unequal this measure will be systematically affected by the sample sizes used in the study. An alternate approach is to adopt a standardizer for computing Cohen's effect size that is not based on a pooled estimate of error variance. The SAS/IML program for the approximate df procedure that was described in Desideratum 8 can be used to compute an effect size that is insensitive to covariance heterogeneity (Keselman et al., 2008). As well, it will compute a confidence interval for an effect size that does not rest on the assumption of a normal distribution of responses; this is accomplished using a bootstrap method.

12. Strengths and Limitations of Repeated-Measures Designs

There are several potential threats to the validity of research findings in a repeated-measures design that should be evaluated in a manuscript. The single-group design lacks a control group for comparison, therefore maturation effects may be impossible to distinguish from time effects. Adopting

a *cohort sequential design* or *step wedge design*, in which the time of entry into the study is staggered, is one approach to estimate these two separate effects.

Within-subject effects might be a result of respondent fatigue, practice effects, or response shift (i.e., a change in the meaning of one's evaluation of the target construct), rather than true change in the outcome. An active area of research in the quality of life literature is around the use of statistical methods, such as SEM, to detect response shift in repeated-measures designs (Schwartz & Sprangers, 1999). A *cross-over design*, in which the provision of treatments is counter-balanced among study participants, can also be used to test for carry-over effects due to fatigue or maturation. Another approach is to externally validate the study results in a different population than the one from which the study sample was selected.

High rates of participant attrition can also threaten the validity of study findings. As noted in Desideratum 6, a sensitivity analysis is one approach to assess potential bias in study parameters as a result of missing data.

Other threats to validity are not unique to repeated-measures designs. Some examples include selection bias due to lack of random assignment to treatment and control groups and experimenter bias, when the individuals who are conducting an experiment have an inadvertent effect on the outcome.

References

- Algina, J., Keselman, H. J., & Penfield, R. D. (2005). Effect sizes and their intervals: The two repeated-measures case. *Educational and Psychological Measurement*, 65, 241–258.
- Arnau, J., Bono, R., Blanca, M. J., & Bendayan, R. (2012). Using the linear mixed model to analyze nonnormal data distributions in longitudinal designs. *Behavior Research Methods*, 44, 1224–1238.
- Arnold, B. F., Hogan, D. R., Colford Jr, J. M., & Hubbard, A. E. (2011). Simulation methods to estimate design power: An overview for applied research. *BMC Medical Research Methodology*, 11, 94.
- Berkovits, I., Hancock, G. R., & Nevitt, J. (2000). Bootstrap resampling approaches for repeated-measures designs: Relative robustness to sphericity and nonnormality violations. *Educational and Psychological Measurement*, 60, 877–892.
- Bird, K. D., & Hadzi-Pavlovic, D. (1983). Simultaneous test procedures and the choice of a test statistic in MANOVA. *Psychological Bulletin*, 93, 167–178.
- Brunner, E., Bathke, A. C., & Placzek, M. (2012). Estimation of Box's ϵ for low- and high-dimensional repeated-measures designs with unequal covariance matrices. *Biometrical Journal*, 54, 301–316.
- Conover, W. J., & Iman, R. L. (1981). Rank transformation as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35, 124–129.
- Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognitive Development*, 11, 121–136.
- Fang, H., Brooks, G. P., Rizzo, M. L., Epsy, K. A., & Barcikowski, R. S. (2008). A Monte Carlo power analysis of traditional repeated-measures and hierarchical multivariate linear models in longitudinal data analysis. *Journal of Modern Applied Statistical Methods*, 7, 101–119.
- Fitzmaurice, G. M., Davidian, M., Verbeke, G., & Molenberghs, G. (2009). *Longitudinal data analysis*. Boca Raton, FL: Chapman & Hall.
- Gomez, E. V., Schaalje G. B., & Fellingham, G. W. (2005). Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics—Simulation and Computation*, 34, 377–392.
- Gribbin, M. J., Chi, Y.-Y., Stewart, P., & Muller, K. E. (2013). Confidence regions for repeated-measures ANOVA power curves based on estimated covariance. *BMC Medical Research Methodology*, 13, 57.
- Guo, Y., Logan, H. L., Glueck, D. H., & Muller, K. E. (2013). Selecting a sample size for studies with repeated-measures. *BMC Medical Research Methodology*, 13, 100.
- Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Bulletin*, 2, 64–78.
- Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: Wiley.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75, 800–802.
- Hoffman, L. (2015). *Modeling within-person fluctuation and change*. New York: Routledge.
- Keselman, H. J. (2005). Multivariate normality tests. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Volume 3, pp. 1373–1379). Chichester: Wiley.
- Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*, 13, 110–129.
- Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophysiology*, 40, 586–596.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

- Liu, S., Rovine, M. J., & Molenaar, P. C. (2012). Selecting a linear mixed model for longitudinal data: Repeated measures analysis of variance, covariance pattern model, and growth curve approaches. *Psychological Methods*, 17, 15–30.
- Lix, L. M., & Keselman, H. J. (1996). Interaction contrasts in repeated-measures designs. *British Journal of Mathematical and Statistical Psychology*, 49, 147–162.
- Lix, L. M., & Lloyd, A. M. (2007). A comparison of procedures for the analysis of multivariate repeated measurements. *Journal of Modern Applied Statistical Methods*, 6, 380–398.
- Looney, S. W. (1995). How to use tests for univariate normality to assess multivariate normality. *The American Statistician*, 49, 64–70.
- Molenberghs, G., Verbeke, G., Demetrio, C. G. B., & Vieira, A. M. C. (2010). A family of generalized linear models for repeated-measures with normal and conjugate random effects. *Statistical Science*, 25, 325–347.
- Newsom, J. T., Jones, R. N., & Hofer, S. M. (2012). *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences*. New York: Routledge.
- Oberfeld, D., & Franke, T. (2013). Evaluating the robustness of repeated-measures analyses: The case of small sample sizes and nonnormal data. *Behavior Research Methods*, 45, 792–812.
- Olson, C. L. (1976). On choosing a test statistic in multivariate analyses of variance. *Psychological Bulletin*, 83, 579–586.
- Roy, S. N. (1958). Step down procedure in multivariate analysis. *Annals of Mathematical Statistics*, 29, 1177–1187.
- Sarkar, S. K., & Chang, C.-K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, 92, 1601–1608.
- Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8, 3–15.
- Schwartz, C. E., & Sprangers, M. A. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality of life research. *Social Science and Medicine*, 48, 1531–1548.
- Shen, S., Beunckens, C., Mallinckrodt, C., & Molenberghs, G. (2006). A local influence sensitivity analysis for incomplete longitudinal depression data. *Journal of Biopharmaceutical Statistics*, 16, 365–384.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.
- Skene, S. S., & Kenward, M. G. (2010a). The analysis of very small samples of repeated measurements II: A modified Box correction. *Statistics in Medicine*, 29, 2838–2856.
- Skene, S. S., & Kenward, M. G. (2010b). The analysis of very small samples of repeated measurements II: An adjusted sandwich estimator. *Statistics in Medicine*, 29, 2825–2837.
- Vallejo, G., Fernandez, M. P., Livacic-Rojas, P. E., & Tuero-Herrero, E. (2011). Selecting the best unbalanced repeated-measures model. *Behavior Research Methods*, 43, 18–36.
- Vallejo, G., Fidalgo, A., & Fernandez, P. (2001). Effects of covariance heterogeneity on three procedures for analyzing multivariate repeated-measures designs. *Multivariate Behavioral Research*, 36, 1–27.
- Verbeke, G., Fieuws, S., Molenberghs, G., & Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, 23, 42–59.
- Westfall, P. H., & Young, S. S. (1993). *Resampling based multiple testing*. New York: Wiley.
- Zhou, M., & Shao, Y. (2014). A powerful test for multivariate normality. *Journal of Applied Statistics*, 41, 351–363.

3

Canonical Correlation Analysis

Xitao Fan and Timothy R. Konold

Pioneered by Hotelling (1935), canonical correlation analysis (CCA) focuses on the relation between two sets of variables, each consisting of two or more variables. In some applications, the two sets may be described in terms of independent and dependent variables, although such designations are not necessary. There are a variety of ways to study relations among groups of variables. The general goal of CCA is to uncover the relational pattern(s) between two sets of variables by investigating how the measured variables in two distinct variable sets combine to form pairs of *canonical variates*, and to understand the nature of the relation(s) between the two sets of variables. CCA has often been conceptualized as a unified approach to many univariate and multivariate parametric statistical testing procedures (Thompson, 1991), and even a unified approach to some nonparametric procedures (Fan, 1996). The close linkage between CCA and other statistical procedures suggests that the association between two sets of variables often needs to be understood in our statistical analyses: “most of the practical problems arising in statistics can be translated, in some form or the other, as the problem of measurement of association between two vector variates \mathbf{X} and \mathbf{Y} ” (Kshirsagar, 1972, p. 281). From this perspective, CCA has been considered as a general representation of the *general linear model* (Thompson, 1984), unless we consider *structural equation modeling* (see Chapters 33 and 34 of this volume) as the most general form of the general linear model that takes measurement error into account (Thompson, 2000). Interested readers are encouraged to consult additional sources for more technical treatments of CCA (Johnson & Wichern, 2002, ch. 10), for more readable explanations and discussions of CCA (Thompson, 1984, 1991), for understanding the linkages between CCA and other statistical techniques (Bagozzi, Fornell, & Larcker, 1981; Fan, 1997; Jorg & Pigorsch, 2013; Kim, Henson, & Gates, 2010; Yan & Budescu, 2009), and for recently proposed methodological extensions of CCA (Heungsun, Jung, Takane, & Woodward, 2012; Ognjen, 2014; Takane & Hwang, 2002; Tenenhaus & Tenenhaus, 2011) including methods for modeling ordinal data (Mishra, 2009). Recommended desiderata for studies involving CCA are presented in Table 3.1 and are discussed in the subsequent sections.

Table 3.1 Desiderata for Canonical Correlation Analysis.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. Substantive research issues are presented and reasons why CCA is an appropriate and rational analytic choice are discussed.	I, M
2. Two natural/logical variable sets, each consisting of two or more variables, are explicitly justified within the context of the substantive research issues.	I, M
3. If possible, path diagrams should be presented to aid readers' understanding of the conceptual canonical model and the various interpretive facets of the canonical analysis.	M
4. Summary statistics for the two sets of measured variables are presented, including sample size and within-set and between-set correlations.	R
5. Canonical correlations and statistical testing of these canonical correlations are presented and discussed.	R
6. Canonical function coefficients (standardized and/or unstandardized) are presented and discussed.	R, D
7. Canonical structure coefficients are presented and discussed.	R, D
8. Based on the pattern and magnitudes of the canonical function and structure coefficients, appropriate interpretations of canonical functions (canonical variates) are provided.	R, D
9. (Optional). Canonical adequacy and redundancy coefficients may be presented and discussed, in light of some known limitations.	R, D
10. Canonical functions (variates) are related back to the substantive research issues.	R, D
11. CCA results are presented clearly to facilitate readers' understanding of CCA findings.	R, D

* I = Introduction; M = Methods; R = Results; D = Discussion.

1. Appropriateness of Canonical Correlation Analysis

Canonical correlation analysis (CCA) is an analytic technique for examining the multivariate relation(s) between two sets of two or more constructs/variables. Through CCA, it is hoped that the multivariate relational pattern between the two sets of variables can be more parsimoniously understood and described. Early in a manuscript, the link between the substantive research issue(s) and CCA as the analytic approach for investigating the substantive issue(s) should be explicated. The early discussion related to the substantive issue(s) should lay the foundation for the later introduction of CCA as a logical/rational analytic choice for investigating the substantive issue(s). In the Methods section of the manuscript, links between the substantive issue(s) and CCA should be more carefully articulated, and the case for CCA as an appropriate analytical choice for the research issue(s) should be made explicit. Oftentimes, the link between the substantive research issue(s) and CCA as the analytic choice is presented or established through discussion of how the two sets of constructs/variables are involved in the substantive research, and how the relational pattern between the two sets of constructs/variables is the focus of the research. For example, Lewis (2007) laid the foundation for CCA as an appropriate analytic choice through discussion in the Introduction of an interest in examining the relation between perception of risk and social norms (first set of variables) and alcohol involvement measures (second set of variables) in a college population.

Canonical correlation analysis has been used to address a wide range of substantive issues in economics (Lima, Resende, & Hasenclever, 2004; Senturk, 2007), education (Ismail & Cheng, 2005; Johnson, 2007), psychology (Aboaja, Duggan, & Park, 2011; Koehn, Pearce, & Morris, 2013), medicine

(Goncalves, Almeida, Estellita, & Samanez, 2013; Heine, Subudhi, & Roach, 2009), and technology (Lee & Lee, 2012). For example, McDermott (1995) examined canonical relations between children's demographic characteristics (age, gender, ethnicity, social class, region, community size, and their interactions) and measures of cognitive ability, academic achievement, and social adjustment. McIntosh, Mulkins, Pardue-Vaughn, Barnes, and Gridley (1992) examined canonical relations between a set of verbal and a set of nonverbal measures of ability; and Dunn, Dunn, and Gotwals (2006) employed CCA in a multivariate validity study for establishing the construct validity of a new measure on sport perfectionism by relating the subscales of the new measure to those of an established measure.

2. Two Logical Sets of Variables

In CCA, two sets of variables are examined with the goal of understanding the multivariate relational pattern between the two sets as more parsimoniously operationalized by the *canonical correlation*. Conceptually, this relation can be described as a bivariate correlation between two 'synthetic' variables, each of which is based on a linear combination of one set of variables involved in the analysis. In CCA, each of two variable sets consists of two or more variables, and the variables within each set should form a natural/logical group. In addition, there should be a reasonable expectation that the two sets of variables are substantively related, and that the relation between the two sets of variables is of potential research interest.

Early in the manuscript, the two sets of variables should be discussed in terms of why the relation between them is of research interest. Furthermore, there should be some indication as to why the variables within each set are included. For example, an industrial psychologist may be interested in understanding how a set of employee satisfaction variables (e.g., career satisfaction, supervisor satisfaction, and financial satisfaction; based on employees' responses to a survey) relates to a set of employees' job characteristics (e.g., variety of tasks required by the position, position responsibility, and position autonomy; based on supervisors' responses for the positions held by each employee). In this situation, the investigator may reason that employees' satisfaction variables and their position characteristics form two logical groups of variables, and that there is a reasonable expectation that the two sets of variables are related in one or more ways. The two sets of variables might have a complicated relational pattern that would not be obvious through simple inspection of the bivariate correlations. Here, CCA may help to uncover the relational pattern via a more parsimonious representation of the association between the two sets.

A possible example of a poorly conceptualized match between two variable sets might involve the pairing of either of the two sets of variables described above, with a set of employees' physical measurements (e.g., measurements of height, waist, and pulse rate). Here, it would be a formidable task to justify why employees' satisfaction variables and their physical measurements would be naturally and logically grouped into two inter-related variable sets. Further, it would be harder to justify the expectation that job satisfaction variables are somehow related to the physical measurements. The author(s) of the manuscript should provide a reasonable justification for the two groups of variables used in CCA in making the case that CCA is an appropriate analytic choice for the issue(s) at hand.

3. Path Diagrams

Depending on the nature of the manuscript, path diagrams may be considered to help readers understand CCA and its major interpretive facets. In practitioner-oriented substantive journals that have little focus on quantitative methods, such path diagrams typically are not needed. However, in more quantitatively-oriented substantive journals, such diagrams can be helpful in aiding readers' understanding of the CCA analytic model and its various interpretive facets.

As an illustrative example, Figure 3.1 presents a model in which one set of observed variables (\mathbf{X}) consists of three variables (x_1, x_2, x_3), and the other set of observed variables (\mathbf{Y}) consists of two variables (y_1, y_2). The single-headed arrows from the observed variables to the unobserved *canonical variates* (X_1^* and Y_1^* , and X_2^* and Y_2^*) denote the presumed direction of influence, and the canonical variates are derived by using *canonical weights* (also called *function coefficients*) to linearly combine the observed variables (x_1, x_2, x_3 , using the a weights, and y_1, y_2 using the b weights). In a manuscript, the *standardized weights* (see Desideratum 6 below about standardized vs. unstandardized function coefficients) may be inserted into the figure so that readers can more easily see the contribution of each observed variable to its canonical variate. The curved double-headed arrow linking the pair of canonical variates represents the canonical correlation. The total number of canonical correlations (i.e., the number of pairs of canonical variates) possible is equal to the number of the observed variables in the smaller of the two sets, though not all of the canonical correlations may be statistically/practically meaningful. In the example illustrated in Figure 3.1, two canonical correlations (R_{C1} and R_{C2}) are possible. Here again, actual canonical correlations values can be placed in the figure. Last, the double-headed arrows, linking the observed variables in each of the two sets, reflect that the correlations among them are taken into account in the derivation of the canonical function coefficients.

Figure 3.1 depicts the *first* pair of canonical variates:

$$\begin{aligned} X_1^* &= a_1(x_1) + a_2(x_2) + a_3(x_3) \\ Y_1^* &= b_1(y_1) + b_2(y_2) \end{aligned}$$

The Pearson product-moment correlation coefficient between these two canonical variates is the first canonical correlation coefficient R_{C1} , which is the maximum of all possible canonical correlation coefficients that can be extracted from the variables. In a similar vein, the second pair of canonical variates (X_2^* and Y_2^*) can be constructed and a second canonical correlation coefficient (R_{C2}) can be obtained, as shown in Figure 3.1. The construction of the second pair of canonical variates is subject to the orthogonality condition: canonical variates in the second pair (X_2^*, Y_2^*) are subject to the constraint that they are not correlated with either of the canonical variates in the previous pair (X_1^*, Y_1^*). In other words, all correlations across pairs (i.e., $r_{X_1^*X_2^*}, r_{Y_1^*Y_2^*}, r_{X_1^*Y_2^*}, r_{X_2^*Y_1^*}$) are zero. If additional pairs of canonical variates can be extracted, as is the case when more observed variables are included in the design, all subsequent pairs of canonical variates are subject to this orthogonality condition relative to all previously extracted canonical pairs.

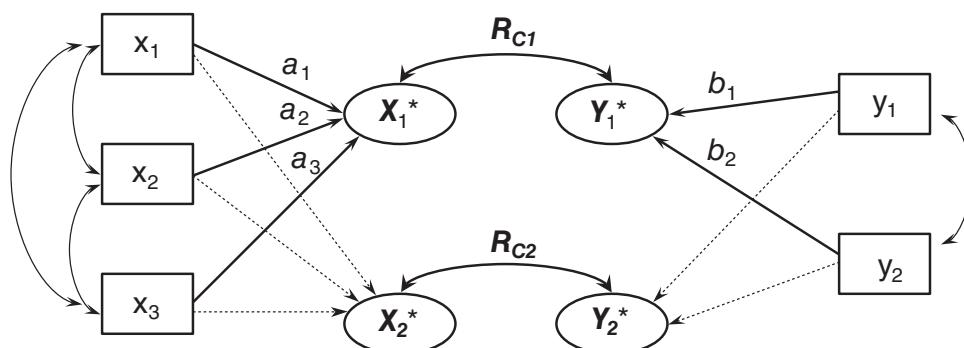


Figure 3.1 Graphic Representation of Canonical Correlation Design.

4. Summary Statistics of Measured Variables

As previously indicated, two sets of logically grouped variables are involved in CCA. It is important that the summary statistics of the two sets of variables are presented in the manuscript. Both within-set and between-set correlations of the two sets of variables should be presented in a form that is easy for readers to see the within-set and between-set relational patterns. Providing such summary statistics serves the general purposes of allowing readers to run secondary data analyses, or allowing readers to replicate the CCA results presented in the manuscript if they have questions regarding the results presented.

As an example of providing such summary statistics, the industrial psychologist mentioned previously is using CCA to investigate how a set of job satisfaction variables (e.g., career satisfaction, supervisor satisfaction, and financial satisfaction; based on employees' responses to a survey) is related to a set of employees' job characteristics (e.g., variety of tasks required by the position, position responsibility, and position autonomy; based on supervisors' responses for each employee on the three aspects of a position). Table 3.2 illustrates how a table of summary statistics for the two sets of variables could be presented. These summary statistics are informative for the readers from the perspective that the correlation pattern is organized into within-set correlation matrices (correlations among variables within each set) and between-set correlation matrix (correlations among variables from different sets: *job satisfaction* variables with *job characteristics* variables).

5. Canonical Correlation Coefficients and Statistical Testing

In CCA, the bivariate Pearson product-moment correlation between two canonical variates within a pair is the canonical correlation coefficient. Unlike the Pearson correlation between two observed variables, however, canonical correlation coefficients do not take on negative values. This is because the direction of a canonical variate [e.g., $Y_1^* = b_1(y_1) + b_2(y_2)$] can be reversed by multiplying all canonical function coefficients (e.g., b_1 and b_2) by -1 . In other words, in the multidimensional space in which the set of multiple variables reside, directionality is arbitrary. For this reason, all canonical correlation coefficients are defined as positive, ranging from 0 to 1.

Like many other statistical techniques (e.g., regression analysis), CCA is a statistical maximization procedure through which the relation between two canonical variates is maximized. The maximization property of the procedure ensures that the first canonical correlation coefficient based on the first pair of canonical variates is the largest among all possible canonical correlation coefficients, and the second canonical correlation coefficient from the second pair of canonical variates is smaller than the previous one, but larger than all *remaining* canonical correlation coefficients, and so on.

Table 3.2 Summary Statistics of Two Sets of Variables ($N=200$).

Job Satisfaction					
Career	1.00				
Supervisor	0.55	1.00			
Finance	0.22	0.21	1.00		
Job Characteristics					
Task Variety	0.31	0.42	0.39	1.00	
Responsibility	0.32	0.42	0.01	0.26	1.00
Autonomy	0.32	0.56	0.37	0.53	0.12
Std	20.41	31.10	2.20	25.77	27.44
					15.82

Note. Variable means have no relevance in CCA, and thus are not presented here.

In CCA, the first question a researcher typically asks is, "Do I have anything?" (Thompson, 2000, p. 301). In other words, the researcher asks if the results appear to indicate that there is some "true" association between the two sets of variables. This question can be addressed from two supplemental perspectives, one being statistical, and the other being practical/substantive.

CCA not only maximizes whatever true population relation may exist between two sets of variables, but it also maximizes any random relation introduced by sampling error. As a result, canonical correlation coefficients can vary in magnitude as a result of sampling error. For this reason, it is necessary to statistically test the canonical correlation coefficients to help ensure that the obtained canonical correlations represent real population relations between the two sets of variables, rather than simply chance relations due to sampling error. Statistically, the question of "Do I have anything?" can be answered by conducting tests of statistical significance for canonical correlation coefficients to determine the probability that the canonical correlation of this or greater magnitude could arise when no "true" relation exists between the X and Y variable sets in the population.

Similar to some other multivariate techniques as explained elsewhere (e.g., Fan, 1997), because of the complexity of sampling distribution theory of canonical correlation coefficients (Johnson & Wichern, 2002; Kshirsagar, 1972) the likelihood ratio test in CCA is a sequential testing procedure rather than a procedure for testing each individual canonical function. For example, consider a CCA that yields three canonical functions, with their respective canonical correlation coefficients. Here, there will be three sequential likelihood ratio tests. The first tests all three canonical functions combined, the second tests the second and third canonical functions combined, and the last tests the third canonical function by itself. Assuming that only the first test is statistically significant, and the latter two are not, this pattern of results leads to the conclusion that the first canonical function is statistically significant, but the latter two are not. Here, our conclusion sounds as though we have conducted significance tests for each individual canonical function, when in reality we have not. Strictly speaking, only the last test in this sequence is a true test for an individual canonical function.

In practice, results based on statistical significance testing should be augmented by measure(s) of effect size (Wilkinson & APA Task Force on Statistical Inference, 1999). Canonical correlation coefficients and their squared values can be taken as a gauge of effect size. In a manuscript, these canonical correlation coefficients, and their squared values, should be clearly presented and discussed. This is particularly true for the statistically significant canonical functions. In general, CCA is a large-sample analytic method and the validity of the likelihood ratio tests for canonical functions depends on a reasonably large sample size. Large sample size leads to high statistical power and, as a result, some trivial canonical association(s) might be statistically significant. Although there is no 'rule of thumb' here, there are some general recommendations. For example, Stevens (2002) recommended the lower limit of sample size of 20 times as many cases as the number of variables for CCA for the purpose of interpreting the first canonical correlation. Barcikowski and Stevens (1975) recommended 40 to 60 times as many cases as the number of variables for two canonical correlations.

Measures of effect size help to temper over-interpreting statistically significant results that might be of low practical importance. As discussed by Cohen (1988), while squared bivariate correlation coefficient (r^2_{xy}) values of 0.02, 0.13, and 0.25 are considered as small, medium, and large effect sizes, respectively, for canonical correlation coefficient the benchmarks for small, medium, and large are dependent on the number of variables in the two sets (X and Y). For example, for a CCA with two variables in each of X and Y sets, squared canonical correlation values of 0.04, 0.24, and 0.45 are considered as small, medium, and large effect sizes, respectively. But for a CCA with four variables in each of the two sets, we need squared canonical correlation coefficient values of 0.06, 0.34, and 0.59 to be considered, respectively, as small, medium, and large. In other words, what is a small or large effect size for canonical correlation coefficient depends on the set sizes used

in the CCA. For details concerning effect size benchmarks for canonical correlation, readers may consult Cohen (1988, ch. 10).

6. Canonical Function Coefficients

Once we ascertain that there are meaningful canonical correlations between the two sets of variables (see Desideratum 5), we proceed to examine the nature of the canonical correlations (“Where do what I have originate?”; Thompson, 2000). For this purpose, function coefficients in CCA will help us to understand the nature of the meaningful canonical correlations.

Similar to other multivariate techniques, CCA produces multiple sets of coefficients. It is important that author(s) provide correct interpretations of these coefficients, including canonical function coefficients. For each of the two sets of variables (e.g., X or Y in the diagram for Desideratum 3), there is a unique set of function coefficients for each canonical function (i.e., canonical variate pair). As a result, there are generally multiple sets of function coefficients, assuming we have more than one canonical function. These canonical function coefficients (i.e., canonical weights) are used to linearly combine the observed variables in each set to obtain a canonical variate for that set and they are derived to optimize the correlation between the pair of canonical variates. They are not necessarily derived to extract the maximum variance from the observed variables. In CCA, the function coefficients serve the primary purpose of determining the canonical variates.

Canonical function coefficients are available in both unstandardized and standardized forms. Both forms represent the partial and unique contribution of an observed variable to its canonical variate, after controlling for the variable’s relation with others in the set. In addition to serving as weights for deriving a canonical variate, these coefficients are often used to gauge individual variables’ relative importance to the resulting canonical variate. Because the variables in a set are often on different measurement scales that will affect the values of unstandardized function coefficients, unstandardized coefficients are generally not useful for assessing the relative importance of the variables. In comparison, standardized coefficients and their associated standardized forms of the variables are placed on the same scale so that they can be more easily compared in terms of their relative contributions. A larger coefficient is interpreted to indicate that a given variable contributes more unique variance to a canonical variate than another variable with a smaller coefficient. In a manuscript, it should be clear that discussions of unique and relative contributions of a variable are based on standardized function coefficients.

Because standardized function coefficients are based on a variables’ partial relation with the canonical variate after the variables’ association with other variables in the set has been removed, the overall variable/variate association might be under- or over-estimated through interpretation of the standardized function coefficients. For example, low standardized function coefficients might underestimate variables’ association with the canonical variate when the variable under consideration is strongly related to both the canonical variate and other variables in the set. In addition, suppression effects can result in sign changes. Because of these issues, function coefficients alone themselves could lead to ambiguity in our understanding about canonical functions.

7. Canonical Structure Coefficients

Canonical structure coefficients provide another mechanism through which linkages between observed variables and their canonical variates can be examined. Unlike CCA function coefficients, which are affected by inter-variable correlations (similar to regression coefficients in regression analysis), a canonical structure coefficient measures the zero-order correlation between a given

variable in a set and a given canonical variate of that set. As such, it reflects the overall degree of association between each variable and the resulting canonical variate that is at the center of the canonical correlation.

There is a general consensus that it is essential that canonical structure coefficients be considered in order to develop good understanding about the canonical functions (e.g., Pedhazur, 1997; Thompson, 2000). Structure and function coefficients of a variable with a canonical variate might be similar, or they might be quite different. When the function and structure coefficients are consistent, either both being low (the variable has little to do with the canonical variate) or both being high (the variable contributes a lot to the canonical variate), it typically does not present any difficulty in interpretation of the results. However, when there is a divergence between function and structure coefficients, some caution is needed in their interpretation. For example, when the function coefficient is low while the structure coefficient is high, the low function coefficient should not lead to the conclusion that the variable shares little with the canonical variate; on the contrary, the variable shares a lot with the canonical variate, but its contribution to the canonical variate overlaps with another or other variables in its own set. In another situation where the function coefficient is moderate or high in absolute value and the structure coefficient is very low, the high function coefficient should not lead to the conclusion that the variable shares a lot with the canonical variate. In this situation, it is very likely that the variable shares little with the canonical variate, and the high function coefficient is the result of a suppression effect, similar to the phenomenon in regression analysis (e.g., Horst, 1941; Lancaster, 1999). These situations, and their implications, are summarized in Figure 3.2 below.

For structure coefficients, values greater than or equal to .32 are often interpreted as practically meaningful. This comes from the fact that squared structure coefficients represent the amount of shared variance between the observed variable and its canonical variate, and that $.32^2$ represents approximately 10% common variance. More importantly, however, a coefficient is usually considered relative to other coefficients in the set, and the relative magnitudes of the coefficients often play an important role in defining a canonical variate. In a manuscript, CCA structure coefficients

		Function Coefficients	
		Low	High
		A	B
Structure Coefficients	Low	The variable shares little with the canonical variate.	A suppression effect is very likely the reason for the moderate/high function coefficient. The variable may actually share little with the canonical variate.
	High	C The variable shares a lot with the canonical variate, but its contribution to the canonical variate overlaps with other variables in its set due to the collinearity between this variable and some other variable(s) in the set.	D The variable shares a lot with the canonical variate. There is little collinearity between the variable and other variable(s) in the set.

Figure 3.2 Relational Patterns of Function and Structure Coefficients.

should be routinely reported and interpreted to help derive an understanding of the nature of the canonical variates.

8. Interpretation of Canonical Functions (Canonical Variates)

Understanding and interpreting the canonical correlation (function) is largely dependent on a meaningful explanation of what the canonical variates represent and these interpretations may be facilitated through applications of orthogonal or oblique rotations (Hironori & Adachi, 2013). The meaning of the canonical variate is usually inferred based on the pattern of coefficients (function and structure coefficients) associated with each of the variables. Similar to loadings in factor analysis (see Chapter 8, this volume), subsets of variables within a set with high and low weights help us to understand the nature of the resulting variate by revealing which variables are most closely associated with it and which variables are not. Here, the researcher is faced with the substantive challenge of understanding the essence of different coefficients with the goal of providing a more parsimonious description of the canonical variate in terms of a construct it is attempting to capture based on the observed variables. For this purpose, variable function coefficients with relatively higher absolute values are given greater emphasis, while lower values are marginalized in the interpretation. The signs of these function coefficients should also be taken into consideration in relation to the scaling of the measured variables. In other words, negative relations between the variables and the canonical variate should be considered and discussed in relation to the substantive labeling of the resulting canonical variate.

As indicated above (see Desideratum 6), when function coefficients are considered, only the variables' unique shared variance with the canonical variate is taken into account. By contrast, structure coefficients consider how a given variable relates to its canonical variate without the interference of how the variable relates to other variables in the set. In this sense, structure coefficients are not confounded by a given variable's relation with other variables in the set. For the purpose of interpreting and labeling the canonical variates, both function and structure coefficients should be considered. In general, when we consider a variable's function and structure coefficients for the purpose of interpreting its canonical variates, the patterns described in both quadrants A and B in Figure 3.2 would suggest that the variable contributes little to the substantive meaning of the canonical variate. On the other hand, the patterns described in both quadrants C and D in Figure 3.2 would suggest that the variable has considerable contribution to the substantive meaning of the canonical variate.

Misinterpretation can easily occur in CCA. For example, unseasoned CCA users might misinterpret the quadrant C pattern (low function coefficient and high structure coefficient) to mean that the variable contributes little to the substantive meaning of the canonical variate, because of the low function coefficient. Similarly, the quadrant B pattern (high function coefficient and very low structure coefficient) can also be misinterpreted to mean that the variable contributes a lot to the substantive meaning of the canonical variate, when in reality the high function coefficient might be the result of a suppression effect. In a manuscript, it should be clear that the interpretation of canonical functions (canonical variates) is based on the joint consideration of both function coefficients and structure coefficients. As Levine (1977) emphasized, if one wants to understand the nature of canonical association beyond the computation of the canonical variate scores (such computation relies solely on function coefficients), one has to interpret structure coefficients.

9. Canonical Adequacy and Redundancy Coefficients

CCA is designed to maximize the canonical correlation (i.e., the correlation between two canonical variates in a pair). It might be of interest to know how much variance a given variate (e.g., \mathbf{X}_1^*

in Figure 3.1) can extract from the variables of its own set (i.e., x_1 , x_2 , and x_3 in Figure 3.1) or how much variance a given variate (e.g., X_1^* in Figure 3.1) can extract from the variables of the other set (i.e., y_1 , y_2 , and y_3 in Figure 3.1). Canonical adequacy and redundancy coefficients, respectively, are designed for such purposes.

Canonical Adequacy Coefficients. Structure coefficients measure the correlation between a variable and its canonical variate, and the squared structure coefficient represents the proportion of variance in the variable that is shared with its canonical variate. *Canonical adequacy coefficients* are associated with each canonical variate and measure the average of all the squared structure coefficients for one set of variables as related to a given canonical variate formed from this set. Canonical adequacy coefficients describe how well a given canonical variate represents the original variables in its set, and quantitatively, it is the proportion of variance in the set of the variables (e.g., x_1 , x_2 , and x_3 in Figure 3.1) that can be reproduced by a canonical variate of its own set (e.g., X_1^*).

Related to the adequacy concept described above, we might also be interested in knowing what proportion of variance in a measured variable is associated with the extracted canonical functions. This percent of variance in a variable associated with the extracted canonical functions is called *communality* (h^2), which is defined in the same way as in factor analysis (see Chapter 8, this volume). If a measured variable has very low communality relative to other variables in a CCA analysis, it suggests that this variable is behaving differently from other variables in the set, and probably does not belong to this set of variables.

Canonical Redundancy Coefficients. Between a pair of canonical variates, the redundancy index (Stewart & Love, 1968) measures the percent of variance of the original variables of one set (e.g., x_1 , x_2 , and x_3 in Figure 3.1) that may be predicted from the canonical variate derived from the other set (e.g., Y_1^*).

Table 3.3 Example of Presentation of CCA Analysis Results.

	Function I			Function II			
	B ^a	r ^b	r ^{2c}	B	r	r ²	h ^{2d}
Job Satisfaction							
x_1 —Career	0.08	0.61	0.37	-0.25	-0.24	0.06	0.43
x_2 —Supervisor	0.83	0.95	0.90	-0.36	-0.30	0.09	0.99
x_3 —Financial	0.32	0.51	0.26	0.98	0.85	0.71	0.97
Adequacy			0.51			0.29	
Rd (R_{X*ys})			0.25			0.03	
R_c		0.70	0.49		0.31	0.10	
Job Characteristics							
Rd (R_{Y*xs})			0.25			0.03	
Adequacy			0.52			0.27	
y_1 —Task Variety	0.24	0.72	0.52	0.74	0.48	0.22	0.74
y_2 —Responsibil.	0.40	0.54	0.29	-0.91	-0.72	0.52	0.81
y_3 —Autonomy	0.69	0.87	0.76	-0.04	0.24	0.06	0.82

a Standardized function coefficients.

b Structure coefficients (when associated with the variables), or canonical correlation coefficients (when not associated with original variables).

c Squared structure coefficients (when associated with the variables); Adequacy and Redundancy coefficients, and squared canonical correlation coefficients.

d Communality: Percent variance in a variable jointly extracted by the two canonical functions.

In other words, when we want to describe how much variance the \mathbf{Y}_1^* variate shares with the set of X variables, or how much variance the \mathbf{X}_1^* variate shares with the set of Y variables (i.e., y_1 , y_2 , and y_3 in Figure 3.1), we use canonical redundancy coefficients. A canonical redundancy coefficient for a canonical variate (e.g., \mathbf{Y}_1^*) can be computed as the product of the canonical adequacy coefficient of its counterpart (i.e., adequacy coefficient for \mathbf{X}_1^*) multiplied by the squared canonical correlation.

There is some controversy surrounding redundancy coefficients in terms of whether they should in fact be interpreted in CCA (Roberts, 1999). The primary concern for redundancy coefficients is that CCA is designed to maximize the canonical correlation, and it does *not* attempt to maximize the redundancy coefficient. As a result, “it is contradictory to routinely employ an analysis that uses function coefficients to optimize [the canonical correlation], and then to interpret results (R_d) not optimized as part of the analysis” (Thompson, 1991, p. 89). Because of this and some other concerns, we caution that the redundancy coefficient is not always meaningful in CCA. Researchers who use this statistic should present appropriate interpretations within their research contexts (e.g., in a multivariate concurrent validity study where there is theoretical expectation for high redundancy coefficients).

10. Substantive Research Issues

CCA provides a variety of interpretive frameworks that should be tied to the substantive question(s) that were outlined early in the manuscript. In general, license for these interpretations comes by way of at least one reliable (i.e., statistically significant) canonical correlation; failure to achieve this suggests that the two sets of variables cannot combine in a way to produce a variate pair correlation that is statistically greater than 0. Assuming the data have passed this threshold, researchers need to relate the resulting canonical functions to the initially proposed research questions. For each statistically significant canonical correlation, the relation between the two variates should be described both in terms of the resulting effect size and in terms of substantive interpretation of the nature of the variate pair in terms of what they represent and why this does or does not align with expectations. As described in previous sections, interpreting and labeling of the variates is best accomplished through the joint use of function and structure coefficients. Although in practice there is a tendency to focus the discussion on which variable(s) contribute most to the variates’ definition and how this relates to the substantive problem, it is also important to consider which variables do not factor into the relations as indicated by very low coefficients.

11. Presentation of CCA Results

As discussed in, for example, Desiderata 6 and 7, CCA typically produces multiple sets of different coefficients which can be quite a challenge for readers to grasp. In a manuscript, it is expected that the major CCA results are adequately presented in some tabular form such that the readers can easily find the major outcomes of the analysis. For each statistically significant canonical function, it is typical to present (1) canonical correlation and/or its squared value, (2) canonical function coefficients, and (3) canonical structure coefficients. In addition, canonical adequacy, canonical redundancy coefficients, and communality for each measured variable may also be presented. For the illustrative data previously presented in Table 3.2, there are three possible canonical functions, with only the first two being statistically significant ($p < .05$). Based on the recommendation of Thompson (2000), Table 3.3 illustrates the major CCA results and how they might be presented in tabular form.

From this presentation, interested readers can more easily find all the relevant CCA statistics, such as the two canonical correlation coefficients (0.70 and 0.31, respectively), function and structure coefficients for the two functions, adequacy and redundancy coefficients associated with each

canonical variate in each of canonical functions, and so forth. Based on the example data presented in Table 3.3, one may tentatively infer that the first canonical function probably represents a general positive relation between job characteristics and employee satisfaction, although Position Autonomy and Satisfaction with the Supervisor appear to play a more important role in defining this first canonical function. The second canonical correlation is primarily defined by the *negative* relation between Financial Satisfaction and Job Responsibility, suggesting that there might be a perception that financial compensation is not in agreement with a position's responsibility.

The presentation in Table 3.3 is succinct and reasonably complete. In a manuscript involving CCA as the major analytic technique, information similar to that presented in Table 3.3 should be expected.

References

- Aboaja, A., Duggan, C., & Park, B. (2011). An exploratory analysis of the NEO-FFI and DSM personality disorders using multivariate canonical correlation. *Personality and Mental Health*, 5, 1–11.
- Bagozzi, R. P., Fornell, C., & Larcker, D. F. (1981). Canonical correlation analysis as a special case of a structural relations model. *Multivariate Behavioral Research*, 16, 437–454.
- Barcikowski, R., & J. P. Stevens. (1975). A Monte Carlo study of the stability of canonical correlations, canonical weights, and canonical variate-variable correlations. *Multivariate Behavioral Research*, 10, 353–364.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Dunn, J. G. H., Dunn, J. C., & Gotwals, J. K. (2006). Establishing construct validity evidence for the Sport Multidimensional Perfectionism Scale. *Psychology of Sport and Exercise*, 7, 57–79.
- Fan, X. (1996). Canonical correlation analysis as a general analytical model. In B. Thompson (Ed.), *Advances in social science methodology* (vol. 4, pp. 71–94). Greenwich, CT: JAI Press.
- Fan, X. (1997). Structural equation modeling and canonical correlation analysis: What do they have in common? *Structural Equation Modeling: A Multidisciplinary Journal*, 4, 65–79.
- Goncalves, A. C., Almeida, R., Estellita, L. M., & Samanez, C. P. (2013). Canonical correlation analysis in the definition of weight restrictions for data envelopment analysis. *Journal of Applied Statistics*, 40, 1032–1043.
- Heine, M., Subudhi, A. W., & Roach, R. C. (2009). Effect of ventilation on cerebral oxygenation during exercise: Insights from canonical correlation. *Respiratory Physiology & Neurobiology*, 166, 125–128.
- Heungsun, H., Jung, K., Takane, Y., & Woodward, T. (2012). Functional multiple-set canonical correlation analysis. *Psychometrika*, 77, 48–64.
- Hironori, S., & Adachi, K. (2013). Oblique rotation in canonical correlation analysis reformulated as maximizing the generalized coefficient of determination. *Psychometrika*, 78, 526–537.
- Horst, P. (1941). The role of predictor variables which are independent of the criterion. *Social Science Research Bulletin*, 48, 431–436.
- Hotelling, H. (1935). The most predictable criterion. *Journal of Educational Psychology*, 26, 139–142.
- Ismail, N. A., & Cheng, A. G. (2005). Analysing education production in Malaysia using canonical correlation analysis. *International Education Journal*, 6, 308–315.
- Johnson, T. E. (2007). Canonical correlation of elementary Spanish speaking English language learners entry characteristics to current English language status. *Education*, 127, 400–409.
- Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Jorg, B., & Pigorsch, U. (2013). A canonical correlation approach for selecting the number of dynamic factors. *Oxford Bulletin of Economics and Statistics*, 75, 23–36.
- Kim, N., Henson, R. K., & Gates, M. S. (2010). Revisiting interpretation of canonical correlation analysis: A tutorial and demonstration of canonical commonality analysis. *Multivariate Behavioral Research*, 45, 702–724.
- Koehn, S., Pearce, A. J., & Morris, T. (2013). The integrated model of sport confidence: A canonical correlation and mediational analysis. *Journal of Sport & Exercise Psychology*, 35, 644.
- Kshirsagar, A. M. (1972). *Multivariate analysis*. New York: Marcel Dekker.
- Lancaster, B. P. (1999). Defining and interpreting suppressor effects: Advantages and limitations. In B. Thompson (Ed.), *Advances in social science methodology* (vol. 5, pp. 139–148). Stamford, CT: JAI Press.
- Lee, J., & Lee, H. (2012). Canonical correlation analysis of online video advertising viewing motivations and access characteristics. *New Media & Society*, 14, 1358–1374.
- Levine, M. S. (1977). *Canonical analysis and factor comparison*. Beverly Hills, CA: Sage.
- Lewis, T. F. (2007). Perceptions of risk and sex-specific social norms in explaining alcohol consumption among college students: Implications for campus interventions. *Journal of College Student Development*, 48, 297–310.
- Lima, M. A. M., Resende, M., & Hasenclever, L. (2004). Skill enhancement efforts and firm performance in the Brazilian chemical industry: An exploratory canonical correlation analysis-research note. *International Journal of Production Economics*, 87, 149–155.
- McDermott, P. A. (1995). Sex, race, class, and other demographics as explanations for children's ability and adjustment: A national appraisal. *Journal of School Psychology*, 33, 75–91.

- McIntosh, D. E., Mulkins, R., Pardue-Vaughn, L., Barnes, L. L., & Gridley, B. E. (1992). The canonical relationship between the Differential Ability Scales upper preschool verbal and nonverbal clusters. *Journal of School Psychology*, 30, 355–361.
- Mishra, S. K. (2009). A note on the ordinal canonical correlation analysis of two sets of ranking scores. *Journal of Quantitative Economics*, New Series, 7, 173–199.
- Ognjen, A. (2014). Discriminative extended canonical correlation analysis for pattern set matching. *Machine Learning*, 94, 353–370.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace College Publishers.
- Roberts, J. K. (1999). Canonical redundancy (Rd) coefficients: They should (almost never) be computed and interpreted. In B. Thompson (Ed.), *Advanced in social science methodology*. (vol. 5, pp. 333–341). Stamford, CT: JAI Press.
- Senturk, A. (2007). A study on the relationship between shares and inflation, exchange rates and interest rates with the use of canonical correlation analysis. *Economic Studies*, 16, 43–64.
- Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NJ: Erlbaum.
- Stewart, D., & Love, W. (1968). A general canonical correlation index. *Psychological Bulletin*, 70, 160–163.
- Takane, Y., & Hwang, H. (2002). Generalized constrained canonical correlation analysis. *Multivariate Behavioral Research*, 37, 163–195.
- Tenenhaus, A., & Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika*, 76, 257–284.
- Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation*. Thousand Oaks, CA: Sage.
- Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development*, 24, 80–95.
- Thompson, B. (2000). Canonical correlation analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 285–316). Washington DC: American Psychological Association.
- Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.
- Yan, H., & Budescu, D. (2009). An extension of dominance analysis to canonical correlation analysis. *Multivariate Behavioral Research*, 44, 688–709.

4

Cluster Analysis

Dena A. Pastor and Monica K. Erbacher

The term *cluster analysis* is generally used to describe a set of numerical techniques for classifying objects into groups based on their values on a set of variables. The intent is to group objects such that objects within the same group have similar values on the set of variables and objects in different groups have dissimilar values. The objects classified into groups are most typically persons and the variables used to classify objects can either be categorical or continuous. Cluster analysis can be used as a data reduction technique to reduce a large number of observations into a smaller number of groups. It can also be used to generate a classification system for objects or to explore the validity of an existing classification scheme. Unlike other multivariate techniques, such as logistic regression or MANOVA (see Chapters 16 and 25, respectively), group membership is not known but instead imposed on the data as a result of applying the technique. Because objects are classified into groups even if no groups truly exist, additional planned analyses beyond cluster analysis are essential. Researchers should use these analyses to provide support for the replicability and validity of a particular cluster solution.

Cluster analytic methods can be classified as being either *model-based* or *non-model-based*. This chapter focuses on non-model-based cluster analytic methods as these are the most commonly used.¹ Model-based clustering methods, such as finite mixture modeling, utilize probability models, whereas non-model-based cluster analytic methods do not utilize statistical models. Thus, non-model-based methods are not formally considered inferential statistics and are more appropriately classified as numerical algorithms.

Popular statistical software packages such as SAS, SPSS, and R can be used to perform non-model-based cluster analysis as well as the replicability and validity analyses. Although a readable overview of cluster analysis can be found in Aldenderfer and Blashfield (1984), this chapter more heavily relies on the more current treatment provided by Everitt, Landau, Leese, and Stahl (2011). Table 4.1 displays the specific desiderata for applied cluster analytic studies, each of which will be explained further below.

1. Objects to Be Classified and Reasons for Classification

Because the purpose of a cluster analytic study is to classify entities into groups, an important question to answer at the forefront of a cluster analytic study is: “What is being classified?” To distinguish among the different types of entities that can be classified using cluster analysis, it is

Table 4.1 Desiderata for Cluster Analysis.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. The objects to be clustered are described and the rationale for classifying these objects into clusters is stated.	I
2. Justification is provided for the specific variables that were chosen as the basis for classifying objects into groups and for those variables that were chosen to provide validity evidence for the cluster solution. If applicable, reliability, and validity evidence for variables is provided.	I, M
3. Any transformation of variables is described and justified.	M
4. Any weighting of variables is described and justified.	I, M
5. The characteristics of the sample are described and a rationale for the sampling method is provided.	I, M
6. Outlying cases and missing data are described. The methods used to address outliers and missing data are explained and justified.	M
7. If utilized, the proximity measure for capturing the similarity or dissimilarity between objects on the set of variables is explicated and justified.	M
8. The specific cluster analytic method used to classify objects into clusters is described in sufficient detail for replication.	M
9. The procedures used for choosing the final cluster solution are explained.	M
10. The methods used to assess the replicability and validity of the final cluster solution are described.	M
11. The software with version number, and specific procedures utilized in the software, are reported.	M
12. Descriptive statistics for all variables are reported.	R
13. Indices or figures used in deciding upon the final cluster solution are provided.	R
14. The final cluster solution is presented, including a description of the characteristics of each cluster.	R
15. Results are provided from the assessment of the solution's replicability and validity.	R
16. The theoretical and/or practical utility of the final cluster solution is addressed and the results are interpreted in light of the research questions and target population.	D
17. Suggestions for future analyses that could be used to provide support for the replicability and validity of the cluster solution are provided.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

useful to consider the $N \times J$ data matrix being used in the analysis, where N equals the number of rows or objects (typically people) and J equals the number of columns or variables. Although cluster analysis is more commonly used to classify N objects into groups based on their values on a set of J variables (i.e., Q analysis), sometimes the technique is used to classify J variables into groups based on their values on a set of N objects (i.e., R analysis). Because cluster analysis can be used to group either objects or variables, it is important to convey which is being classified early in the manuscript.

In social science research, it is rare to see cluster analysis used to classify variables. Researchers often use other techniques, such as factor analysis (Chapter 8, this volume), for this purpose. If cluster analysis is used to group variables, it is important for researchers to justify its use over these more common alternatives.

The more common use of cluster analysis is to classify objects, typically individual people. However, objects other than people (e.g., countries, schools, stimuli, or words) can serve as the rows in a data matrix and thus as the entities to be grouped. It is important to convey precisely the nature of the objects being classified, even if those objects are people.

After clarifying what objects are being classified, the manuscript should provide an answer to the question: “What goals are to be accomplished by classification?” Cluster analysis may be pursued for the purposes of data reduction, to generate a classification scheme, or to validate an existing classification scheme. Readers are directed to Romesburg (1984) for a more thorough list of possible goals in a cluster analytic study. Answering the question of why objects need to be classified is an important piece of a cluster analytic manuscript since it justifies the use of the technique and also provides a framework under which the utility of the results can be gauged.

2. Variables Used to Classify Objects into Groups and to Provide Validity Evidence

Variables used to classify objects into groups (internal variables). In cluster analysis, objects are assigned to clusters based on their values on a set of variables, often called internal variables or the cluster variate. Because the particular variables chosen by the researcher drive the resulting classification scheme, readers should be provided with answers to the following questions: “Why were the internal variables selected?” and “What manifestations of the variables were used?”

The authors need to address why the internal variables are thought to be essential for separating objects into groups. The justification for the use of certain variables over others should be linked to theory, previous research and the goals the authors hope to attain by classification. Cluster solutions are highly dependent upon the variables used; the inclusion of even a single irrelevant variable can complicate the discovery of true clusters (Milligan, 1980). Thus, this justification is critical. If a numerical variable selection procedure (Steinley & Brusco, 2008) is used to differentiate useful from irrelevant variables, a description of the procedure should be provided along with a rationale for its use.

Given the dependency of the cluster solution on the internal variables, it is imperative the variables are described in detail in the Methods section. Different classification schemes may result if a different manifestation of a variable is utilized (e.g., test anxiety measured using self-report vs. galvanic skin response). When tests or scales are used to measure variables, supporting reliability and validity evidence should be provided (see Chapter 29, this volume).

If it is not obvious, the level of measurement (categorical, continuous) of the variables should be reported. Although reporting the level of measurement is not typical, it is important in this context because the choice of proximity measure (see Desideratum 7) depends on whether the set of internal variables are categorical, continuous, or a mix of both metrics.

Variables used to provide validity evidence (external variables). Justifying the variables chosen to provide validity evidence, or external variables, for the final cluster solution is also important. Cluster analysis can result in a final solution with more than one cluster, even when no unknown groups exist in the population. Providing validity evidence is the first step in supporting the viability of the final cluster solution as anything other than an artifact of the analysis. Authors should answer the following questions: “What variables should be related to the final cluster solution and how should they be related?” and “What theories dictate these relationships?”

External variables are not restricted by measurement properties. They can be nominal, ordered categorical, continuous, composite scores, or anything else, as long as the appropriate analyses are used (see Desiderata 10 and 15). The level of measurement of external variables should be reported, as well as evidence of reliability and validity for any tests or scales measuring these variables.

The identification of a theoretical backing for external variables is key. Authors should identify one or more theories supporting the use of each variable. These theories provide the foundation for evaluating the validity evidence the author will compile after finding a final cluster solution.

3. Transformation

Many sources advise internal variables be transformed prior to their use in a cluster analysis (e.g., to z -scores). Because there is a wide variety of transformations available, and because the cluster solutions resulting from untransformed versus transformed variables can differ, answers should be supplied to the following questions: “Were variables transformed and if so, why and how were they transformed?”

When the variables used to classify objects are on different scales (e.g., SAT scores versus grade point average), transforming these internal variables is often recommended to avoid variables with larger variances overpowering the classification of objects into groups. Although the cluster analysis literature often advocates for the transformation of variables, authors need to carefully consider whether transformation is necessary, which depends (in part) on whether the clustering method (see Desideratum 8) is invariant to transformations of the data. For instance, some clustering methods will yield the same solution regardless of whether variables are raw or transformed (e.g., non-hierarchical methods minimizing the determinant of the within-cluster dispersion matrix, \mathbf{W}).

Any transformation of the variables, or lack thereof, should therefore be reported and justified. Because a wide variety of transformations exist (Milligan & Cooper, 1998; Steinley, 2004a), authors need to describe and provide a rationale for the transformation method employed. Justification is particularly essential for the z -score transformation due to its consistently poor performance in simulation studies (Milligan & Cooper, 1998; Steinley, 2004a). For this reason, authors may want to consider other transformations, such as standardization by range or by some estimate of within-group variance. Authors are also encouraged to consult and reference simulation studies examining the performance of various transformation methods under different conditions (e.g., Gnanadesikan, Kettenring, & Tsao, 1995; Milligan & Cooper, 1988; Schaffer & Green, 1998; Steinley, 2004a; Steinley & Brusco, 2008).

4. Weighting

When internal variables are equally weighted, each variable influences the resulting classification scheme to the same degree. When variables are differentially weighted, some variables have a stronger influence on the resulting classification scheme than others. Sometimes differential weighting is explicit and intended. Other times, differential weighting occurs implicitly through the use of variables on different scales, heterogeneous variances, strong variable relations or the transformation method employed. The manuscript should thus address the question: “Were the variables differentially weighted?”

Sometimes weighting of internal variables is done explicitly, to heighten the influence certain variables have on the resulting classification scheme. When variables are weighted explicitly, the particular weighting method should be described in detail and justification for variable weights should be provided. The differential weighting of internal variables is related to discussion in Desideratum 2 pertaining to the selection of variables. Just as justification needs to be provided for excluding variables from the cluster variate (essentially giving such variables a weight of zero), it also needs to be provided in the Introduction section for included variables weighted more heavily than others. In other words, the authors must convey why the more heavily weighted internal variables are allowed to have a stronger influence on the classification of objects into clusters.

Differential weighting of internal variables that occurs implicitly, as opposed to explicitly, should also be addressed by the authors. Recall from Desideratum 3 that untransformed variables with larger scales (SAT vs. grade point average) or with larger variances may more heavily influence the

cluster solution. If untransformed variables are utilized, the differential weighting of variables that may result should be addressed. Authors should also acknowledge any effects the transformation method, if used, has on variable weighting (see Desideratum 3).

Implicit weighting of variables can also occur when internal variables are highly related to one another. To explain, consider two variables with a correlation of 1.00. These variables have 100% overlapping variance. If both are included in the cluster variate, it is like including either of the variables twice. The overlapping variance is counted twice, or given too much weight. If there are strong relations among variables, the authors should explain how they identified such relations (e.g., bivariate correlations) and how they chose to handle the multicollinearity. Some options include using Mahalanobis distance as the proximity measure, excluding redundant variables, or representing related variables as composites or factors using principal components analysis or factor analysis, respectively (see Chapter 8, this volume). Note if the latter approach is taken, solutions using composites or factors may differ in undesirable ways from those using the original variables (Chang, 1983; Green & Krieger, 1995; Schaffer & Green, 1998).

5. Sample

The information provided in this desideratum is intended to assist the authors in answering the following questions: “What particular sample of objects was used in the study and why?” and “How were data collected?” Even though non-model-based cluster analytic techniques do not qualify as being inferential statistical procedures *per se*, care still needs to be placed on sample selection as the cluster solution is highly dependent on the objects used in the analysis. In less exploratory applications of cluster analysis, where the number and nature of the groups have been suggested by theory or previous research, extant literature suggests objects be sampled randomly from a target population in order to facilitate generalization (Everitt et al., 2011; Milligan, 1996; Milligan & Hirtle, 2012). However, because non-model-based cluster analysis is non-inferential, it is important to keep in mind that “any conclusions the researcher ascribes to the larger population from which the sample was obtained must be based on analogy, not on inferential statistics” (Romesburg, 1984, p. 30).

If a particular cluster solution is anticipated, objects should be selected that represent the unknown groups. Milligan (1996) and Milligan and Hirtle (2012) note it is not essential to sample objects in proportion to their cluster size, but do suggest oversampling objects anticipated to be in small clusters to help ensure their recovery in the cluster solution.

Multiple samples should always be sought to explore the replicability of cluster solutions (see Desideratum 10). As is typical in most manuscripts, the Methods section should include a detailed description of the sample, including the sample selection method, sample size and characteristics. This information allows readers to judge the adequacy of the sample and the generalizability of the results.

6. Outlying Cases and Missing Data

Other questions that need to be answered about the data include: “Were there any outlying cases or missing data and if so, what was their nature and how were they handled?” Outlying cases (objects) can overly influence the results of a cluster analysis (Milligan, 1980), with different techniques being more or less sensitive to the presence of outliers (Belbin, Faith, & Milligan, 1992; Milligan, 1989). It is recommended that the Methods section contain an explanation of the methods by which outliers were identified (e.g., histograms, scatterplots, Mahalanobis distance), including any outlier detection technique specific to cluster analysis (e.g., Cerioli, 1998; Cheng & Milligan, 1996; Cuesta-Albertos, Gordaliza, & Matran, 1997). Additionally, the number and nature of identified outlying cases should

be stated and treatment of these cases described. Similarly, authors should summarize the nature and extent of missing data. Methods used to treat missing data, such as listwise deletion or multiple imputation (Silva, Bacelar-Nicolau, & Saporta, 2002), should be explicitly stated and justified.

7. Proximity Measures

Proximity measures, which may or may not be used by a clustering method, capture the similarity or dissimilarity between two objects on the set of internal variables. Because a wide variety of proximity measures exist and because cluster solutions resulting from the use of different proximity measures can differ, the authors should answer the following question: “Was a proximity measure used, and if so, which proximity measure was used?”

Some cluster analytic methods do not use proximity measures, instead they operate directly on the $N \times J$ data matrix. In these situations, there is no need to report a proximity measure. The majority of clustering methods (see Desideratum 8), however, utilize proximity measures. Proximity measures can be obtained directly, for example by acquiring ratings of the extent to which two objects are similar, or indirectly through calculation of measures that utilize the variables’ values for object pairs. Regardless of how they are obtained, the result is an $N \times N$ matrix of such measures, where N is the number of objects. When a proximity measure is used, authors should describe it in sufficient detail and provide justification for the particular proximity measure chosen.

If proximity measures are calculated for binary variables, the authors should state how the proximity measure handles joint absences. For instance, consider a binary variable that is coded 0 to represent the absence of a characteristic and 1 to represent the presence of a characteristic. Objects that share an absence of the characteristic are considered to be joint absences and objects that share a presence of the characteristic are considered joint presences. It is important to consider whether joint absences should be weighted the same as joint presences. Proximity measures for binary variables differ not only in how they weight joint absences and presences, but also in how they weight matches (0/0, 1/1) and non-matches (0/1). Thus, it is important to describe the proximity measure and justify its weighting scheme.

Explanation should also be provided for how proximity measures were calculated for any ordinal or nominal variables. If internal variables with mixed levels of measurement are utilized, the authors need to explicate how the different metrics were handled when creating the proximity matrix.

Proximity measures cannot only be classified by the kind of variables for which they are best suited, but also by the kind of clusters they are most inclined to capture. A proximity measure should be chosen that corresponds well with the unknown clusters’ characteristics. An educated guess at these characteristics could be provided by theory, previous research, or preliminary visualizations of the data (see Everitt et al., 2011, ch. 2). A manuscript would be strengthened by including a discussion of the methods used to anticipate the clusters’ characteristics and the ability of the proximity measure to capture such characteristics.

8. Cluster Analytic Method

One of the most important decisions made in cluster analysis is the method chosen to separate objects into clusters. This choice should be guided by a method’s ability to effectively recover the types of clusters anticipated, even in the presence of noise or outliers (Everitt et al., 2011; Milligan, 1996; Milligan & Hirtle, 2012). Authors are thus encouraged to describe the characteristics of anticipated clusters (number, size, shape) and refer to simulation studies indicating the chosen method’s ability to recover clusters with similar characteristics, even in less than ideal conditions. It is important for authors to demonstrate that their choice was not based on convenience or popularity, but

on careful review of research investigating performance of various methods and consideration of their clusters' anticipated characteristics. In the absence of any knowledge regarding cluster characteristics, it is recommended that authors choose a method (and, if applicable, proximity measures) suited to uncover clusters with a variety of characteristics.

Exclusive vs. Fuzzy. There is an overwhelming number of cluster analytic methods from which to choose. Deciding whether an exclusive (also called hard or crisp clustering) or fuzzy clustering technique is needed can help limit the possibilities. Exclusive clustering methods constrain each object to belong to one and only one cluster. Fuzzy clustering methods allow each object to belong to all clusters and quantify the strength of belonging with a degree of membership statistic or a membership weight (Tan, Steinbach, & Kumar, 2006; Xu & Wunsch, 2005). Typically, weights for a single object sum to one across all clusters. In the social sciences, exclusive methods are more conventional, with fuzzy clustering methods more commonly seen in neuroscience research, computer science, and related fields. Thus, this chapter focuses on exclusive clustering methods. However, resources on a wide variety of exclusive and fuzzy clustering procedures (Xu & Wunsch, 2005), their performance in terms of validity coefficients (Jafar & Sivakumar, 2013), and comparisons among fuzzy clustering algorithms on a variety of desirable properties (Baraldi & Blonda, 1999) are readily available in the literature.

Given the wide range of cluster analytic methods available, it is important for authors to report whether they are using exclusive versus fuzzy clustering and to provide a rationale for their choice. If an exclusive cluster analytic method is used, authors should clarify whether the method is hierarchical or non-hierarchical. Information that should be reported for hierarchical and non-hierarchical methods is reviewed in turn below.

Hierarchical. With hierarchical methods, no set number of clusters is specified a priori. Instead, N partitions are made of the data, either beginning with all objects in their own cluster and progressing to a single cluster (*agglomerative*) or beginning with all objects in a single cluster and progressing until each object is in its own cluster (*divisive*). Whether the author is using an agglomerative or divisive partitioning should be made explicit. Although some researchers are interested in how objects and clusters are combined at each step, it is more common to choose one solution from the N solutions provided (see Desiderata 9).

All hierarchical methods result in nested clusters, where solutions with more clusters are nested within solutions with less clusters. If hierarchical methods are chosen, the authors need to provide a rationale for using a method that yields a nested classification of the data.

Hierarchical clustering methods differ in the rules (i.e., which are functions of the proximity measures) used to combine or divide clusters. The different rules are often reflected in the method's name. Popular hierarchical methods include: single linkage, complete linkage, average linkage, centroid linkage, median linkage, and Ward's method. Authors should identify the name of the hierarchical method (and any alternative names), describe the method, and provide justification for its use.

Non-hierarchical. The majority of non-hierarchical methods require the user to specify the number of clusters before implementing the method. It is typical for this number to remain fixed during implementation of the procedure, although some non-hierarchical procedures allow the number of clusters to change (see Steinley, 2006). It is therefore important for users to report if the employed non-hierarchical method required the number of clusters to be specified a priori and if this number was fixed during implementation. Authors should report all numbers of clusters implemented and ideally provide justification for the number or range of numbers chosen (e.g., "we suspected at least two groups and thus examined solutions with two to six clusters").

Objects are assigned to one of k clusters in non-hierarchical methods for the purposes of maximizing or minimizing some numerical clustering criterion. For instance, the goal of the k -means

method, which is a popular non-hierarchical approach, is to assign objects to clusters so that the distance of objects from the center of their cluster (i.e., the cluster centroid) is a minimum. Other methods use different clustering criteria (e.g., minimizing the determinant of \mathbf{W} , the within-cluster dispersion matrix). Because different methods use different clustering criteria, and because naming conventions for methods are not always consistent in the literature, it is essential that authors report and describe the criterion being employed.

To ascertain the ‘best’ allocation of objects to clusters, the clustering criterion could be calculated for all possible partitions of N objects into k clusters, with the solution having the lowest (or highest) value of the criterion serving as the optimal solution. Because this approach is computationally demanding, optimization algorithms are used, which begin with an initial allocation of objects to clusters and then iteratively reallocate objects until no or only trivial improvement is seen in the clustering criterion between iterations.

When using non-hierarchical methods, it is important to report how objects were initially allocated to clusters due to the vast number of ways this can be accomplished. For instance, the partitioning from a hierarchical clustering method can be used as the initial classification or objects can be randomly assigned to partitions. Alternatively, cluster centroids (cluster seeds) can be randomly generated or supplied by the user. Because different initialization methods lead to different solutions (Blashfield, 1976; Friedman & Rubin, 1967), the initialization method utilized must be described.

Simulation studies have indicated several non-hierarchical methods are prone to producing locally optimal solutions (Steinley, 2003). Locally optimal solutions are those that prematurely terminate and thus fail to truly maximize or minimize the clustering criterion. Encountering local optima is somewhat dependent on the quality of the initialization method used. Using informative seed values (e.g., centroids from a hierarchical solution or based on previous research and theory) can help, but does not guarantee local optima will not be encountered (Steinley, 2003). If only one k -cluster solution is obtained, the authors should acknowledge the possibility their solution represents a local optimum. Running the k -cluster solution multiple times using different seed values would strengthen a manuscript. This approach allows researchers not only to investigate the extent to which locally optimal solutions might be occurring, but also to select the ‘best’ solution.

It is also important to describe how objects were reallocated to clusters during the procedure as this can differ across methods. For instance, updates to the clustering criterion (or to cluster centroids in k -means) may happen after the reallocation of each object, after a fixed number of reallocations, or after all objects have been reallocated. This further emphasizes the need to describe the non-hierarchical clustering method in detail. Simply supplying the name of the procedure (e.g., k -means) does not provide readers with enough detail to replicate the study or judge the adequacy of the clustering methods used.

9. Choosing a Final Cluster Solution

Cluster analysis is often criticized because of the subjectivity associated with choosing the number of clusters to retain; therefore, it is essential an answer be provided to the following question: “What information was used to choose the final cluster solution?”

Because any cluster analysis technique will return a partitioning of the data, even in the absence of true clusters, it is important for researchers to consider the possibility that no clusters exist in their data (i.e., $k = 1$). At the very least, this possibility should be acknowledged in the manuscript. At the most, the results of a formal hypothesis test could be provided (Steinley & Brusco, 2011). If evidence supports the existence of clusters, authors must address how many clusters exist. It is critical to describe the procedures used to choose the number of clusters. Ideally, authors will use a

variety of methods, with agreement across methods strengthening this decision. In addition to the plots and indices (described below) that are typically used to choose a cluster solution, authors are strongly encouraged to consider the interpretability of the solution along with its correspondence to theory and past research.

Hierarchical. Commonly used procedures to select the final solution include inspection of the *dendrogram* (a tree-like plot showing the progression of objects being merged into clusters) or inspection of a graph plotting the number of clusters by some form of the average within cluster variance (often referred to as the fusion or amalgamation coefficient). Because the interpretation of such plots is highly subjective, statistical rules have been created to assist with choosing a cluster solution. For instance, a simulation study by Milligan and Cooper (1985) examined the performance of 30 different statistical indices in detecting the correct number of clusters. Milligan (1996) suggested two or three of the better performing indices be utilized in hierarchical cluster analysis, with agreement among indices providing stronger evidence for a particular solution.

Non-hierarchical. Although the number of clusters, k , is often specified a priori in non-hierarchical clustering methods, it is strongly encouraged that solutions with a variety of values for k be examined. Thus, one should avoid examining a single solution of k clusters and instead explore a variety of non-hierarchical solutions. The exact values of k examined should be made explicit in the manuscript. When more than one solution is examined, a flattening of the function of the final clustering criterion plotted against k can be used to choose among solutions. As with the plots described for hierarchical methods, the subjectivity of this method is considered problematic. The indices examined by Milligan and Cooper (1985) can be adapted for use with non-hierarchical methods, so long as the index is not limited to use with only nested clusters (Steinley, 2006). The use of such indices with non-hierarchical methods and the development of new indices for determining k in k -means clustering is an active area of research (Steinley, 2006, 2007; Steinley & Brusco, 2011). A brief description of any indices used, whether old or new, should be provided along with their associated references.

10. Replicability and Validity

Even if no structure exists in the data, clusters will be created as a result of implementing the clustering method. In fact, clustering methods are often criticized because they may create structure, regardless of whether structure truly exists. To address this criticism, it is essential authors show that the classification scheme resulting from a cluster analysis is consistent and meaningful. Thus, a cluster analytic study is not complete until it answers the following: “Is the final cluster solution replicable?” and “Is the final cluster solution meaningful?”

Replicability. The replicability of a cluster solution refers to the consistency or the stability of the clustering method across different samples. Ideally, two samples would be collected for this purpose, although in practice it may be more feasible to randomly split a single sample into two halves. If sample size permits, replicability could be examined by independently implementing the clustering method in each sample and exploring the extent to which the same cluster solution emerges in both samples. More formal procedures for replication, similar to cross-validation techniques, are described by McIntyre and Blashfield (1980) and Breckenridge (1989, 2000). Results of replication analyses should not be used to decide on the number of clusters to retain because a cluster solution can replicate even in data with no clustering structure (Steinley & Brusco, 2011).

Validity. If a solution replicates across samples, more faith can be placed in the generalizability and meaningfulness of the results. Although replicability is necessary for a solution’s validity, evidence beyond replicability is needed for the solution to be considered meaningful, authentic, and valid.

Validity evidence for a cluster solution is acquired by showing the clusters relate to external variables in ways anticipated by theory, logic and previous research. Validity thus involves embedding the solution in a program of construct validity.

The variables used to provide validity evidence for the cluster solution are called ‘external’ because they are not the same variables used to create the clusters. Although it may seem appealing to illustrate the meaningfulness of the solution by showing clusters significantly differ in their ‘internal’ variable values, it is not noteworthy since the clusters were created to be maximally different from one another on these variables. Differences among clusters on internal variables is useful for describing the cluster solution (see Desideratum 14), but is not considered an appropriate approach for conveying the validity of a solution.

Because the validity of a solution is essential to a cluster analytic study, the external variables need to be carefully selected. A description of the external variables and justification for their use in assessing the validity should be provided early in the manuscript (see Desideratum 2). The Methods section should contain a detailed description of the external variables as well as the analyses used to assess relationships between the clusters and external variables. For reporting purposes, readers are encouraged to refer to the chapters in this book corresponding to the statistical methods chosen for their validity analyses.

11. Software

Blashfield (1977) found when different software packages are used, different solutions can be obtained even if the same clustering method is applied to the same data. Although this finding may not hold as true today, it underscores the need to provide answers to the following questions: “What is the name and version of the software program used?” and “If applicable, which specific procedure was used within that software program?”

A variety of software programs can be used to conduct cluster analysis (e.g., Clustan, SAS, SPSS, R). In the Methods section the author should report the name and version of the software program used and, if applicable, the specific procedure (e.g., proc cluster, proc fastclus, or proc modeclus within SAS). If R is used, the reference for the version of R should be provided, along with any packages used to conduct cluster analysis (e.g., pvclust, NbClust) and their respective references.

12. Descriptive Statistics

A common requirement of almost all quantitative research is the reporting of descriptive statistics. Descriptive statistics for all internal and external variables should be reported or made available. Descriptive statistics provide information as to the level of measurement of the variables, enabling readers to judge the appropriateness of proximity measures (if used) and validity analyses. The descriptive statistics for the internal variables are particularly useful because they convey information about the variables for the overall sample, which may be of interest when the same information is reported by cluster (see Desideratum 14). Measures of association between all internal variables should also be provided to assess multicollinearity (see Desideratum 4).

13. Indices and Figures

In Desideratum 8, authors were instructed to describe the figures and/or indices used in choosing the final cluster solution. Such figures or indices should be displayed in the Results section or made available to the reader upon request. It is recommended to report indices for the range of

solutions considered (not just the final solution) so readers can judge for themselves the solution to be favored. The extent to which indices or figures were in agreement should also be discussed. If multiple solutions were considered, authors should describe the process used to arrive at the final solution. Essentially, all information used by the authors to choose the final solution should be reported, which should include the interpretability of the solution and its correspondence with previous research and theory.

14. Final Cluster Solution

The results section should clearly state which cluster solution was chosen and contain a thorough description of the resulting clusters. The clusters can be described by providing a table with the proportion of the sample in each cluster and relevant descriptive statistics for the variables by cluster. It may also be useful to provide a figure to illustrate the cluster characteristics. For instance, a graph of the variable means by cluster may nicely convey the cluster profiles. Other graphs are described by Everitt et al. (2011, §9.6).

It is recommended clusters be referred to by a letter or number (Cluster 1, Cluster 2 or Cluster A, Cluster B) rather than by a name. Although this may result in seemingly uninformative names for the clusters, it is beneficial in that it allows for a more objective interpretation of the clusters and helps to avoid misleading names that may bias readers' perceptions and interpretations of the results.

15. Replicability and Validity Results

In Desideratum 9 it was recommended authors provide the methods used to assess the replicability and validity of the final cluster solution. In the Results section, the authors need to present the results of the replicability and validity analyses in enough detail to allow the reader to judge for themselves the consistency and meaningfulness of the final solution.

With regard to replicability, if the cluster analysis was conducted independently on different samples, the extent to which the same solution was championed across samples should be reported and the characteristics of the clusters described (see Desideratum 14). If more of a cross-validation type of analysis was conducted, the associated statistics for the analysis should be reported (e.g., adjusted Rand index; Hubert & Arabie, 1985; Steinley, 2004b).

A wide variety of statistical analyses can be pursued to examine the validity of a cluster solution. When reporting results of these analyses, readers should consult the chapters in this book that correspond to the statistical analyses used.

16. Utility of Final Cluster Solution

A thoughtful examination of the study's results should be provided in the discussion section. This examination should provide answers to the following questions: "Did the final cluster solution correspond to the clusters anticipated based on previous research or theory?" and "Do the authors consider the results of the replicability and validity analyses to be supportive?" Possible explanations should be provided for any unanticipated results, including any results that do not support the replicability and validity of the final cluster solution. If a meaningful cluster solution was obtained with evidence supporting its replicability and validity, then a discussion should ensue as to whether the goals underlying object classification were met. This discussion ties the results of the study back to the purposes stated in the introduction (see Desideratum 1). The utility of the cluster solution also needs to be addressed. In other words, the ways in which the classification scheme can be used in

future research or in applied settings should be explicated. Additionally, the generalizability of the final cluster solution should be addressed, for example by reminding the reader of the population(s) from which the data was sampled. In particular, if the authors did not replicate the solution across multiple samples, they should acknowledge here that evidence is needed to support the generalizability of the cluster solution.

17. Future Analyses

It is a well-known fact in measurement that the collection of reliability and validity evidence for an instrument or scale is a never-ending process. The same can be said about cluster analysis. Once a solution is obtained, evidence must be continually gathered to support the consistency and authenticity of the solution. If a seemingly stable and meaningful cluster solution is found, the authors should provide next steps in establishing the replicability and validity of the solution. Suggestions may include the conditions under which replicability should be further assessed or other external variables to include in future validity studies.

Note

1 Although this chapter focuses on non-model-based clustering methods, many of the desiderata in Table 4.1 also apply to model-based clustering methods. Desiderata generally pertaining to model-based approaches are 1–2, 5–6, 8–17. Because desiderata will differ for model-based approaches, reporting guidelines for model-based classification approaches should be consulted (e.g., the guidelines provided in this volume for latent class analysis, Chapter 12, and latent variable mixture models, Chapter 15). Readers interested in learning about model-based methods are referred to Everitt, Landau, Leese, and Stahl (2011), Everitt and Hand (1981), Banfield and Raftery (1993), Fraley and Raftery (1998, 2002), and McLachlan and Peel (2000).

References

- Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Newbury Park, CA: Sage Publications.
- Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803–821.
- Baraldi, A., & Blonda, P. (1999). A survey of fuzzy clustering algorithms for pattern recognition. I. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 29(6), 778–785.
- Belbin, L., Faith, D. P., & Milligan, G. W. (1992). A comparison of two approaches to beta-flexible clustering. *Multivariate Behavioral Research*, 27, 417–433.
- Blashfield, R. K. (1976). Mixture model tests of cluster analysis. Accuracy of four agglomerative methods. *Psychological Bulletin*, 83, 377–385.
- Blashfield, R. K. (1977). The equivalence of three statistical packages for performing hierarchical cluster analysis. *Psychometrika*, 42, 429–431.
- Breckenridge, J. N. (1989). Replicating cluster analysis: Method, consistency, and validity. *Multivariate Behavioral Research*, 24, 147–161.
- Breckenridge, J. N. (2000). Validating cluster analysis: Consistent replication and symmetry. *Multivariate Behavioral Research*, 35, 261–285.
- Cerioli, A. (1998). A new method for detecting influential observations in nonhierarchical cluster analysis. In A. Rizzi, M. Vichi, & H. H. Bock (Eds.), *Advances in data science and classification* (pp. 15–20). Berlin: Springer.
- Chang, W. C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, 32, 267–275.
- Cheng, R., & Milligan, G. W. (1996). K-means clustering methods with influence detection. *Educational and Psychological Measurement*, 56, 833–838.
- Cuesta-Albertos, J. A., Gordaliza, A., & Matran, C. (1997). Trimmed K-means: An attempt to robustify quantizers. *Annals of Statistics*, 25, 553–576.
- Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions*. London: Chapman & Hall.
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Chichester: John Wiley & Sons.
- Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41, 578–588.
- Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97, 611–631.
- Friedman, J. H., & Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159–1178.

- Gnanadesikan, R., Kettenring, J. R., & Tsao, S. L. (1995). Weighting and selection of variables for cluster analysis. *Journal of Classification*, 12, 113–136.
- Green, P. E., & Krieger, A. M. (1995). Alternative approaches to cluster-based market segmentation. *Journal of the Market Research Society*, 37, 221–239.
- Hubert, L. J., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193–218.
- Jafar, O. A. M., & Sivakumar, R. (2013). A comparative study of hard and fuzzy data clustering algorithms with cluster validity indices. In *Proceedings of International Conference on Emerging Research in Computing, Information, Communication, and Applications* (pp. 775–782). New York: Elsevier Publications. Retrieved from http://searchdl.org/public/book_series/elsevierst/1/120.pdf.
- McIntyre, R. M., & Blashfield, R. K. (1980). A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research*, 15, 225–238.
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
- Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45, 325–342.
- Milligan, G. W. (1989). A study of the beta-flexible clustering method. *Multivariate Behavioral Research*, 24, 163–176.
- Milligan, G. W. (1996). Clustering validation: Results and implications for applied analyses. In P. Arabie, L. J. Hubert, & G. De Soete (Eds.), *Clustering and classification* (pp. 341–375). River Edge, NJ: World Scientific Publication.
- Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, 5, 181–204.
- Milligan, G. W., & Hirtle, S. C. (2012). Clustering and classification methods. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (2nd ed., pp. 189–210). Charlottesville, VA: John Wiley & Sons.
- Romesburg, C. H. (1984). *Cluster analysis for researchers*. Belmont, CA: Lifetime Learning Publications.
- Schaffer, C. M., & Green, P. E. (1998). Cluster-based market segmentation: Some further comparisons of alternative approaches. *Journal of the Market Research Society*, 40, 155–163.
- Silva, A. L., Bacelar-Nicolau, H., & Saporta, G. (2002). Missing data in hierarchical classification of variables—a simulation study. In K. Jajuga, A. Sokolowski, H.-H. Bock (Eds.), *Classification, clustering, and data analysis: Recent advances and applications* (pp. 121–128). Berlin: Springer.
- Steinley, D. (2003). K-means clustering: What you don't know may hurt you. *Psychological Methods*, 8, 294–304.
- Steinley, D. (2004a). Standardizing variables in K-means clustering. In D. Banks, L. House, F. R. McMorris, P. Arabie, & W. Gaul (Eds.), *Classification, clustering and data mining applications* (pp. 53–60). New York: Springer.
- Steinley, D. (2004b). Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods*, 9, 386–396.
- Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59, 1–34.
- Steinley, D. (2007). Validating clusters with the lower bound for sum of squares error. *Psychometrika*, 72, 93–106.
- Steinley, D., & Brusco, M. J. (2008). Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika*, 73, 125–144.
- Steinley, D., & Brusco, M. J. (2011). Choosing the number of clusters in k-means clustering. *Psychological Methods*, 16, 285–297.
- Tan, P., Steinbach, M., & Kumar, V. (2006). Cluster analysis: Basic concepts and algorithms. In *Introduction to data mining*. Boston, MA: Addison-Wesley.
- Xu, R., & Wunsch, D. C. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16, 645–678.

5

Correlation and Other Measures of Association

Jill L. Adelson, Jason W. Osborne, and Brittany F. Crawford

Correlation and, more generally, association is a basic staple of inferential statistics, often easily computed with handheld calculators, simple spreadsheet software, and even by hand. Most modern researchers compute correlations and other indices of relatedness without thinking deeply about their choices. Despite being basic in the current pantheon of analytic options, measures of association are important, almost ubiquitous, and can easily produce errors of inference if not utilized thoughtfully and with attention to detail.

The goal of correlational analyses is to assess whether two variables of interest covary or are related, and ultimately, to draw conclusions that allow the researcher to speak to some issue in the “real world.” In the best of instances, the researcher has (a) a theoretical rationale for exploring this issue, (b) high-quality measurement of variables of interest, (c) an appropriate analytical approach, (d) attention to detail in terms of ensuring assumptions of the analytic approach are met, and (e) followed best practices in performing the analysis and interpreting the result(s).

The primary focus of this chapter will be the most common index of association, Pearson’s product–moment correlation coefficient, r , but also will include brief discussions of other measures of relation, such as alternative correlation coefficients and odds ratios. Related, but more advanced, correlational procedures such as multiple regression, logistic regression, multilevel modeling, and structural equation modeling are beyond the scope of this chapter (but see Chapters 16, 22, 23, and 33, this volume). Of course, as modern statistics incorporates these and all ANOVA-type analyses into a single general linear model, many of our comments will apply across a broad range of techniques. Treatments of issues in this chapter are included in classic and more recent textbooks such as Cohen, Cohen, West, and Aiken (2003), Pedhazur (1997), Aiken and West (2010), Tabachnick and Fidell (2014) and Osborne (2017), among others. Specific desiderata for studies using correlation/relational methodologies are presented in Table 5.1 and explicated subsequently.

1. Substantive Theories and Measures of Association

“Fishing expeditions,” such as examining large correlation matrices to look for ideas to explore, are really a sub-optimal and limiting way to go about trying to understand the world. Rather, good

Table 5.1 Desiderata for Correlation and Other Measures of Association.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. A succinct literature review situates the current study in the field's context. The substantive theory or rationale that led to the investigated relation(s) is explained.	I
2. The goals and the correlational nature of the research question(s) are clearly stated.	I
3. Hypotheses acknowledge a linear relation or curvilinear relationship and any possible interactions.	I
4. The variables of interest are explicitly identified and operationalized.	I, M
5. Relevant psychometric characteristics (preferably of the current data but possibly from previous, similar studies) are presented and discussed. At minimum this should include reliability and factor structure (if applicable). Variables with unacceptable reliability or validity evidence should not be included in analyses.	M, R
6. The sampling framework and sampling method(s) are defined and justified.	M
7. Results from (preferably <i>a priori</i>) power analyses that are in line with the chosen sampling strategy are reported.	M
8. If data to be analyzed are nested/multilevel in nature, or otherwise more appropriate for multilevel modeling, those methods are used or standard errors are adjusted appropriately, depending on the research question.	M, R
9. Fundamental descriptive statistics of the variables are presented and discussed (e.g., measurement scale, mean, variance/standard deviation, skewness, and kurtosis).	M, R
10. If preliminary analyses suggest that data on variables of interest are not reasonably normally distributed, appropriate actions are taken to normalize the data or subsequent analytic strategies that accommodate significant deviations from normality are chosen (and justified as appropriate).	M, R
11. Restriction of range and homogeneous populations are examined through preliminary analyses and addressed if appropriate.	M, R
12. Missing data, if present, are appropriately dealt with.	M, R
13. Authors report how outliers/fringeliers were defined, identified, and, if any were present, how they were dealt with.	M, R
14. Tests of assumptions underlying the analyses are presented.	M, R
15. The appropriate correlational test (Pearson's <i>r</i> , tetrachoric correlations, etc.) is used, dependent on the measurement scale of the variables and the likelihood of meeting distributional assumptions.	R
16. Multiple zero-order analyses are accompanied by mention of, and possibly corrections for, increased Type I error rates. Multiple analyses are combined into fewer multivariate analyses when justified.	R
17. Authors use semipartial and partial correlations where appropriate and interpret them correctly.	R, D
18. Where appropriate, an interpretation of results takes variable transformations into account.	R, D
19. <i>p</i> values are interpreted correctly and accompanied by confidence intervals. Once statistically significant effects are found, effect sizes, not <i>p</i> values, should guide the narrative and discussion.	R, D
20. Appropriate effect size measures are reported and interpreted.	R, D
21. Curvilinear relations or interactions, when found, should be presented graphically.	R, D
22. Discussion of correlational analyses exercises appropriate caution regarding causal inferences.	R, D

* I = Introduction, M = Methods, R = Results, D = Discussion.

research flows from good theory. Thus, reviewers should evaluate any type of quantitative analysis from within the knowledge base of the discipline that anchors the research.

Of course, data can inform theory just as easily as theory can drive research. Yet, unless we clearly ground the current research being reported in the context of what has come before and is currently being discussed (where appropriate), we risk merely reinventing the wheel. No research should occur without clearly articulating how previous knowledge has led to the specific questions addressed within the study at hand.

Thus, reviewers should see a thorough review of the literature that has led to the current study, and, if citations only go back a decade or two, reviewers should make sure authors are not missing seminal research in the field. Few lines of research have roots less than 10–20 years old.

2. Goals and Correlational Nature of the Research Question(s)

It is worth addressing early on in this chapter that the term *correlational* can be used both in reference to research design and an analysis technique. In reference to a type of design, the term correlational indicates that the variables in a study were not manipulated, only measured to describe the relation between the two variables. A correlational analysis technique refers to the process of determining whether a relation is present between two variables, and, if so, measuring the strength and direction of this relation. The remainder of this chapter uses the term correlational in reference to a statistical analysis technique.

Reviewers should demand clearly stated, operationalizable goals and objectives that clearly lend themselves to correlational/relational analyses. All too often, one sees goals for group differences, growth, or change inappropriately explored through correlational methods (as well as correlational hypotheses tested via ANOVA-type designs). Researchers should be encouraged to align their analytic strategies to their goals/data as closely as possible, and reviewers need to enforce best practices. In the case of correlational analyses, the goals and objectives should be clearly focused on the relational nature of two or more variables. Words such as “group difference” or “causal” should not be included in correlational hypotheses. This being said, correlational analyses can be done using experimental data or in the context of theoretical models that include proper control.

3. Hypotheses Regarding the Nature of the Relation

As human nature is complex and interesting, it is likely that many associations are not linear in nature. Consider the well-known association between arousal (or less accurately, anxiety) and performance, such that increased arousal leads to increased performance. However, researchers also have noted that there is an optimal level of arousal, and beyond that, higher levels tend to be associated with diminishing performance until, at some theoretical level of extreme arousal, performance would be as low as extremely low arousal. In fact, the classic arousal-performance curve discussed here will produce correlations close to $r = 0.00$ if curvilinearity is not accounted for, leaving a robust and important effect undetected or misestimated. Thus, it is in the interest of the researcher, as well as the scientific community, to consider carefully whether the relations that they plan to examine are linear or would exhibit a curvilinear relation or interaction effect. Hypotheses should acknowledge whether the posited relation is theorized to be linear or not.

4. Operationalizing the Variables of Interest

Researchers interested in the social sciences often intuitively understand the need to explicitly identify and describe the variables of interest; at the same time, they are challenged to operationalize

important but almost unmeasurable constructs. One can have the most thorough and impressive literature review, sound theory and rationale, and important goals, but without being able to *validly and reliably* measure the constructs at hand, resulting conclusions are just as erroneous, if not more, than those which could be drawn from a thought experiment. Operationalization (defining specifically how one did or will measure the constructs at hand) is the *conditio sine qua non* of science, but this is often the place in the logic of the research project where the scientific quality of the project breaks down. Reviewers must demand high quality operationalization in order to get high quality research outcomes.

Because we are dealing with variable relations in this chapter, we must know (a) exactly what variables the researchers were examining for a potential association, (b) how the authors defined each considered construct, (c) what *specific* operations the researchers utilized to measure each variable, and (d) how successfully the researchers measured what they thought they measured. For example, some researchers have studied whether students' use of instructional technology is related to student achievement. But what, exactly, do they mean by "instructional technology" or "student achievement"? In the modern context, researchers have referred to instructional technology as a catch-all phrase for computers, the Internet, personal data assistants (PDAs), calculators, smartboards, student response systems, online assessment systems, word processors, multimedia learning systems, and so forth. And, of course, student achievement can refer to many different things, from high-stakes end-of-grade tests mandated in many states to ephemeral concepts such as change in knowledge or acquisition of skills. One could imagine many different studies purporting to test whether instructional technology is related to student achievement, all using radically different operationalizations of each variable, each potentially coming to divergent conclusions at least partly because of a different operational definition rather than a flaw in the basic premise. Thus, the reader should have a precise picture of exactly what was meant by instructional technology and both how and how well student use of instructional technology was assessed. The last point refers to the quality of measurement, which will be discussed below.

Finally, it is at this point that the reviewer makes an assessment as to whether the basic measurement was adequate or not. For example, if a researcher is assessing student use of instructional technology, and the researcher merely asks students to indicate on a survey whether they used the Internet:

- (a) every day,
- (b) twice a week or more,
- (c) at least once a month, or
- (d) less than once a month.

Has the researcher measured this construct with high precision and quality? Not at all. Leaving aside the inherent issues with self-report data, the question—and potential response—is imprecise (we do not know *why* students were using the Internet or what they did with it). Yet, the researcher could easily correlate resulting data with scores on a high-stakes test. Drawing conclusions for any other constructs beyond the one originally measured may lead to the publishing of erroneous conclusions. It is important that reviewers disallow poorly operationalized research from becoming part of the knowledge base of the field, as it can skew results dramatically.

5. Psychometric Characteristics

In a recent survey of the educational psychology literature in top-tier journals, only about one in four studies (26%) reported any type of basic psychometric information on the measures utilized in

the research being reported (Osborne, 2008b). Because reliable and valid measurement is a hallmark of good science and a necessary condition for *any* statistical analysis such as correlation or regression, authors should report basic psychometric information on their measures, such as internal consistency and/or other forms of reliability and validity (see Chapter 29, this volume). Further, reviewers should ensure that reliability of measurement and evidence of validity of inferences is acceptable in magnitude.

Acceptable standards for reliability have been debated for decades, and often depend on the type of reliability being reported. Many authors assume that internal consistency estimates (Cronbach's alpha) of .70 and above are acceptable (e.g., Nunnally, 1978), and despite the fact that social science journals are filled with alphas of .70 and below, this should not be considered quality measurement. In fact, an alpha of .70 represents measurement that is approximately 50% error variance, and even when reliability is .80, effect sizes are attenuated dramatically (when alphas are in the .80 to .85 range, correlations can be attenuated approximately 33%; Osborne, 2003). Of course, internal consistency estimates are not always appropriate or ideal (e.g., reliability of behavioral observation data), but reviewers should have *some* reasonable indicator of reliability and validity reported in every paper, and the reviewers should be convinced that the data are sufficiently reliable to be taken seriously.

Where scales are used to measure constructs, authors should be encouraged to present evidence that the scale structure they are endorsing for further analysis is the best, most parsimonious representation of that scale via confirmatory factor analysis (see Chapter 8, this volume) or other modern measurement analysis (e.g., Rasch measurement or IRT; see Chapter 11, this volume).¹ Factor structure matters because if authors are creating composites of items that are not measuring a homogenous construct, they are introducing error into their measurement, which significantly attenuates effect size estimates and power to detect effects.

In sum, authors should research their instruments and, when possible, choose those that tend to produce highly reliable data and allow for valid inferences. If authors are dealing with data that have less than optimal reliability, simple analyses such as correlation and regression can be manually disattenuated via common formulae (Cohen et al., 2003; Pedhazur, 1997) or through other, more sophisticated means such as structural equation modeling (see Chapter 33, this volume).

6. Justification of Sampling Method(s)

There are many studies with interesting conceptual frameworks but flawed sampling methods. For example, the social science literature contains studies purporting to study working adults but using adolescent psychology students; or investigations intending to generalize results to school-age children but utilizing college students; or studies of the US population as a whole (with all the racial/ethnic, chronological, religious, and developmental diversity) but using samples that are quite homogeneous in nature.

Strong inference requires good sampling. The sample must be representative of the population of interest in order for the results to generalize. Studies should clearly describe the intended population and the methods used to recruit and retain participants for the study that are representative of that population, along with response rates, and so forth. The results should confirm that the sampling strategy was successful by, where possible, showing that sample characteristics are not meaningfully different from known population characteristics.

The sampling strategy should be chosen to sample the population of interest. It is not always the case that we want a sample that is purely representative of the population as a whole. Sometimes a researcher will want to compare different subgroups, and where subgroups are not equally common within a population, a stratified sampling framework should be used,² the rationale for

how the strata were chosen and sampled should be defended, and the analyses and discussion of the results should take it into account. For example, if a researcher wanted to explore the relation between income and life satisfaction across the lifespan, that researcher should be careful to use stratified sampling to gather equal numbers of people at various ages because random sampling from the population will not yield equal numbers of individuals of each age range (see also Chapter 35, this volume).

7. Power Analysis

Statistical power is the probability of rejecting a null hypothesis when indeed it is false given a particular effect size, alpha level, sample size, and analytic strategy.³ Jacob Cohen (e.g., Cohen, 1962, 1988) spent many years encouraging the use of power analysis in planning research, reporting research, and interpreting results (particularly where null hypotheses are not rejected). Authors were discussing the issue of power more than 60 years ago (e.g., Deemer, 1947), yet few authors in the social sciences today (only about 2% in Educational Psychology) reported having tested or used power (Osborne, 2008b). The concept of power is complementary to significance testing and effect size, and, Cohen and others argued, a necessary piece of information in interpreting research results (see Chapters 6 and 27, this volume).

Null Hypothesis Statistical Testing (NHST) has been reviled in the literature by many as counter-productive and misunderstood (e.g., Fidler & Cumming, 2008; Thompson, 2002). The ubiquitous $p < .05$ criterion is probably not disappearing from scholarly social science any time soon despite the fact that almost all null hypotheses are ultimately false in an absolute sense (e.g., rarely is a population correlation coefficient exactly 0.00; rarely are two population means exactly equal to all decimal places), and thus, given sufficient power, even the most minuscule effect can produce a $p < .05$ (see also Tukey, 1991). Thus, the reporting of effect sizes (in the case of simple correlation, r and r^2 are the effect sizes to report), which tells us generally how important an effect is, is crucial.

Power is critical in two different aspects of research. First, Cohen and others argued that no prudent researcher should conduct research without first making *a priori* analyses to determine the probability of correctly rejecting the null hypothesis. Researchers who fail to do this risk committing Type II errors (failing to reject a false null hypothesis), thus wasting the time and effort conducting underpowered research and worse, risk causing confusion in the field by publishing conflicting results that may not be conflicting at all, merely a collection of studies that would have all been in accord had all been sufficiently powered. Additionally, researchers who fail to do *a priori* power analyses risk gathering too much data to test their hypotheses—if a power analysis indicates that data from $n = 100$ participants would be sufficient to detect a particular effect, gathering a sample of $n = 400$ is a waste of resources. Second, power analyses are useful in order to shed light on null results. For example, when a researcher fails to reject a null hypothesis with an *a priori* power of .90 to detect anticipated effects, s/he can be fairly confident of having made a correct decision. However, when a null hypothesis is not rejected, but there is low power, it is unclear as to whether a Type II error has occurred.

Further, something not generally discussed in the social science literature is that low power has implications for Type I error rates in bodies of literature. In the ideal case of strong power (i.e., almost all real effects are detected) and the almost ubiquitous alpha of .05 (i.e., few false conclusions of effects where there are none), a relatively small proportion of studies achieving statistical significance in a field would result from Type I errors. However, maintaining an alpha of .05 but considering a sub-optimal situation where a large group of studies have low power (e.g., .20), in fact a much larger relative proportion of published studies would contain Type I errors because so few true effects are being detected (Rossi, 1990). Thus, ironically, poor power in studies can inflate the

proportion of published studies with significant effects that are based on Type I errors. This may, in turn, lead to seemingly conflicting results in a line of research, and give rise to apparent controversies in fields that traditionally have poor power to detect effects which are, in the end, more likely a result of poor power than conflict of a substantive nature.

In sum, statistical power is an important concept, but authors and reviewers tend to neglect this piece of the empirical puzzle. Authors should report *a priori* (preferably) or *a posteriori* power to detect effects, and reviewers should insist on seeing it.

8. Nested Data

People tend to exist within organizational structures, such as families, schools, business organizations, churches, towns, states, and countries. In education, students exist within hierarchical social structures that can include families, peer groups, classrooms, grade levels, schools, school districts, states, and countries. Workers exist within production or skill units, businesses, and sectors of the economy, as well as geographic regions. Health care workers and patients exist within households and families, medical practices and facilities (e.g., a doctor's practice, or hospital), counties, states, and countries. Many other communities exhibit hierarchical data structures as well.

In addition, Raudenbush and Bryk (2002) discussed two other types of data hierarchies that are less obvious: repeated-measures data and meta-analytic data. Once one begins looking for hierarchies in data, it becomes obvious that data repeatedly gathered on an individual are hierarchical, as all the observations are nested within individuals (who are often nested within other organizing structures). Although there are other adequate procedures for dealing with repeated-measures data, the assumptions relating to them are rigorous (see Chapter 2, this volume) and the procedures relating to hierarchical linear modeling require fewer assumptions (see Chapter 22, this volume). Also, when researchers are engaged in the task of meta-analysis—the analysis of a large number of existing studies (see Chapter 19, this volume)—it should become clear that participants, results, procedures, and experimenters are nested within experiment or study.

Hierarchical, or nested, data present several problems for analysis. First, individuals that exist within hierarchies tend to be more similar to each other than people randomly sampled from the entire population. For example, students in a particular third-grade classroom are more similar to each other than to students randomly sampled from the population of third graders. This is because students are not randomly assigned to classrooms from the population but rather are assigned to schools based on geographic factors. Thus, students within a particular classroom tend to come from a community or community segment that is more homogeneous in terms of morals and values, family background, socio-economic status, race or ethnicity, religion, and even educational preparation than the population as a whole. Further, students within a particular classroom share the experience of being in the same environment—the same teacher, physical environment, and similar experiences—which may lead to increased homogeneity over time.

The problem of independence of observations. The previous discussion indicated that, often, study participants tend to share certain characteristics (e.g., environment, background, experience, demographics) and hence their data are not fully independent. However, most analytic techniques require independence of observations as a primary assumption. Because this assumption is violated in the presence of hierarchical data, ordinary least squares (OLS) regression produces standard errors that are too small (unless these *design effects* are incorporated into the analysis; see McCoach & Adelson, 2010, for a discussion of dealing with dependence via design effect and multilevel modeling,⁴ as well as Chapter 35 of this volume). When assumptions of independence are not met, an undesirable outcome is that the smaller standard errors bias significance testing toward inappropriate rejection of null hypotheses.

The problem of how to deal with cross-level data. Going back to the example of a third-grade classroom, it is often the case that a researcher is interested in understanding how environmental variables (e.g., teaching style, teacher behaviors, class size, class composition, district policies or funding, or state or national variables) are related to individual outcomes (e.g., achievement, attitudes, retention). But given that outcomes are gathered at the individual level whereas other variables are assessed at the classroom, school, district, state, or nation levels, the question arises as to what the unit of analysis should be and how one should deal with the cross-level nature of the data.

One way researchers traditionally attempt to deal with such data is to assign classroom or teacher (or school or district) characteristics to all students, bringing the higher-level variables down to the student level. The problem with this approach is non-independence of observations as all students within a particular classroom assume identical scores on a variable.

Another way researchers have attempted to deal with such data has been to aggregate data up to the level of the classroom, school, or district levels. Thus, one could assess the effect of teacher or classroom characteristics on *average* classroom achievement. However, this traditional approach is problematic in that: (a) much (up to 80–90%) of the individual variability on the outcome variable is lost, potentially leading to dramatic under- or over-estimation of observed relations between variables (Raudenbush & Bryk, 2002), and (b) the outcome variable changes significantly and substantively from individual achievement to average classroom achievement.

Neither of these conventional and common strategies for dealing with multilevel data can be considered best practice. Neither allows the researcher to examine truly important and interesting questions, such as the relation of a particular teacher characteristic with student learning. The most common way to deal with nested data in the social and behavioral sciences is through multilevel modeling. Both aggregation and disaggregation can lead to wildly misestimated effects, with some effects (particularly in aggregated analyses) overestimated by 100% or more (see Osborne, 2008c). Thus, nested data not analyzed with multilevel modeling, or some other design-based corrective action (see, e.g., Chapter 35 of this volume or Lee, Forthofer, & Lorimor, 1989), should be treated as suspect and might be rejected by reviewers.

Further, performing multilevel analyses such as hierarchical linear modeling has no drawbacks if model assumptions are met. If observations are not independent the results are correct estimates of effects accounting for non-independence, protecting the researcher from an error of inference or substantially misestimated effects. If observations are truly independent, multilevel analyses will exactly reproduce the results of simple correlation or OLS regression analyses.

9. Descriptive Statistics

The presentation and discussion of descriptive statistics is important because it facilitates meta-analyses of studies—a best practice for a variety of reasons (see Chapter 19, this volume)—and re-analyses of data can be accomplished more easily if basic statistics are reported. Further, descriptive information helps readers (and authors) understand the data on which the main analyses are based.

10. Normality of Data

Normally distributed data are desirable in most cases, even when utilizing nonparametric techniques. Few authors report normality statistics for their data, and fewer still report correcting for non-normality despite the fact that it is often relatively easy to do, involving simple mathematical manipulation or removal of outliers. It should be noted that misapplications of these transformations can make matters worse, rather than better, so authors and reviewers should not only assess whether data are acceptably

normal (e.g., skew between -1.0 and 1.0 , closer to 0 is better), but should deal with non-normality appropriately. For example, it is often the case that data transformations are most effective if the minimum value of a distribution is anchored at a value of 1.0 exactly (e.g., Osborne, 2008a). Further, for most simple, common transformations to work, original data need to be positively skewed. For distributions with a negative skew, the distribution needs to be *reflected* or reversed by multiplying each data point by -1.0 and adding a constant to bring the minimum value to 1.0 . Correlations and most other common statistical analyses benefit significantly from these practices, making the results more generalizable. When they are done, authors must be careful to interpret results appropriately and to acknowledge to what population(s) the results generalize if outliers were removed.

11. Restriction of Range

Both homogeneous groups and ceiling or floor effects (i.e., restriction of range) cause attenuation problems for correlations. With groups that are homogeneous, or very similar on the variables of interest, the relation between the variables cannot be accurately detected. For instance, if the researcher was interested in the relation between mathematics achievement and enjoyment of mathematics in middle school students but only sampled students in a summer enrichment program for mathematically talented students, the data would have a restriction of range; that is, the sample would only represent one small section of the plot in the population as these students would not have much variance in mathematics achievement or enjoyment. Therefore, the correlation would be attenuated.

Ceiling and floor effects, in which the instrument does not measure high or low enough, respectively, are particularly problematic in education research in which achievement tests may be too easy or too difficult for the group being tested and also may be problematic in psychological research in which instruments do not allow for enough variability in the sample. Not only does this create very little variance (similar to homogeneous groups), but the relation between the variables may even look as if it is not linear. For instance, when examining the relation between IQ and achievement on a grade level achievement test, assuming a linear relation, students at a particular level of IQ begin hitting the ceiling of the achievement test. Although the relation may have been strong and positive up to that point, the scatter begins to bend as IQ continues to increase but the scores on the achievement test cannot continue to increase as students already are achieving the highest level. Thus, the correlation coefficient would be attenuated, and the scatter would not be linear. Some statistical methods, such as the Tobit model, are appropriate for data that exhibit a ceiling effect.

12. Missing Data

Missing data are common in the social sciences, yet perusal of the literature shows that few authors report whether there were any missing data, how missing data issues were addressed, and perhaps equally important, whether participants with missing data were statistically significantly different in some way than those with complete data (i.e., whether the missing data were randomly distributed). Many authors use listwise deletion, simple removal of participants' data with any missing score on any variable in the analysis. However, this approach is sub-optimal and can lead to significant changes in the nature of the data and decreases in generalizability. Another popular approach, mean substitution, can have unintended consequences (such as artificially decreasing the standard deviation of the variable) and should be avoided. Multiple imputation, a Monte Carlo technique in which missing data are replaced with plausible values from a distribution not once but $m > 1$ times, is a better alternative than either of the aforementioned options (e.g., Cole, 2008). Irrespective, reviewers should expect to see a discussion and justification of exactly how authors dealt with missing data, if present.

13. Outliers and Influential Data Points

Although there has been (and continues to be) great discussion in the methodology literature regarding the definition of outliers and whether or not they have significant effects on effect estimates (summarized in Osborne & Overbay, 2004), empirical examples of their effects are common. Outliers can have significant effects on the accuracy of effect estimates. Osborne and Overbay (2004) demonstrated that even in reasonably large samples of up to 400 or more subjects, a handful of outliers with scores that are only slightly outside the distribution (z scores of 3.00 to 3.50) can cause substantial errors in parameter estimates and in inference. Across hundreds of simulations, summarized below in Table 5.2, 70–100% of correlation estimates were statistically significantly more accurate after identification and removal of fringeliers, and errors of inference were reduced dramatically (note also that errors of inference were disturbingly prevalent with just a few outliers present; Osborne (2013, 2017) shows how models can be evaluated for these issues).

Thus, reviewers should expect authors to report checking for outliers and fringeliers, and report how they were handled. There are several ways authors can handle extreme scores such as these, including deleting them, recoding them to some more reasonable value, selectively weighting cases to reduce their influence on outcomes, or using data transformations to reduce their influence.

14. Underlying Assumptions

Many authors are under the erroneous impression that most statistical procedures are “robust” to violations of most assumptions. Writers such as Zimmerman (1998) have pointed out that violations of assumptions can lead to serious consequences, and when violations of more than one assumption are present, it is not safe to assume the validity of the analysis results.

Authors routinely fail to report testing assumptions (only 8.3% of top tier educational psychology articles reported testing assumptions in 1998–1999; Osborne, 2008b), which could indicate that authors fail to test assumptions or authors fail to report associated results. In either case, this is a serious issue as the quality of the corpus of knowledge depends on high quality research analyses, and failing to test assumptions can lead to serious errors of inference and misestimation of effects.

For example, failing to use multilevel modeling or design-based adjustment to test relations when data are nested (independence of error terms or independence of observations assumption) can

Table 5.2 Effects of Outliers on Correlations.

<i>Population r</i>	<i>N</i>	Average initial <i>r</i>	Average cleaned <i>r</i>	<i>t</i>	% more accurate	% errors before cleaning	% errors after cleaning	<i>t</i>
<i>r</i> = -.06	52	.01	-.08	2.5**	95%	78%	8%	13.40***
	104	-.54	-.06	75.44***	100%	100%	6%	39.38***
	416	0	-.06	16.09***	70%	0%	21%	5.13***
<i>r</i> = .46	52	.27	.52	8.1***	89%	53%	0%	10.57***
	104	.15	.50	26.78***	90%	73%	0%	16.36***
	416	.30	.50	54.77***	95%	0%	0%	—

Note: 100 samples were drawn for each row. Outliers were actual members of the population who scored at least $z = 3$ on the relevant variable.

With $N = 52$, a correlation of .274 is significant at $p < .05$. With $N = 104$, a correlation of .196 is significant at $p < .05$. With $N = 416$, a correlation of .098 is significant at $p < .05$, two-tailed.

** $p < .01$, *** $p < .001$.

Source: Osborne and Overbay (2004)

cause substantial misestimation of effects (see Desideratum 8). Failure to measure variables reliably (see Desideratum 5) can lead to serious underestimation of effects in simple (zero-order) relations and either under- or overestimation of effects when variables are controlled for depending on the relative reliability of the covariates and variables of interest (e.g., Osborne & Waters, 2002). As Osborne and Waters (2002) pointed out, failure to test for and account for other issues such as curvilinearity (assumption of linear relationship) can lead to serious errors of inference and underestimation of effects. In sum, while authors are ultimately responsible for reporting results from testing underlying assumptions, reviewers and editors are equally culpable when such results do not appear in published research reports. Reviewers should demand that authors report results from testing each assumption of a utilized statistical method.

Before reporting Pearson's r , specifically, authors must check that they have not violated the necessary assumptions, many of which are easily overlooked. These assumptions include: (a) both variables are continuous (at the interval or ratio levels of measurement); (b) linearity (the relationship between the variables can be displayed by a straight line on a scatterplot); (c) there are no significant outliers or the outliers have been detected and removed; (d) the variables are bivariate normally distributed; (e) homoscedasticity (the variances of the predicted values along the line of best fit are comparable for all values of the predictor variable). Violated assumptions should be addressed individually and corrected when possible, otherwise, alternative measures should be considered (see Osborne, 2017, Ch.2 for detailed information about these assumptions).

15. Pearson's r Alternatives

In many textbooks, authors prominently discuss alternative correlation coefficients (e.g., Cohen et al., 2003, pp. 29–31) such as the *point-biserial* correlation (a simplified Pearson's r appropriate for when one variable is continuous and one is dichotomous), *phi* (ϕ) (a simplified Pearson's r when both variables are dichotomous), and *Spearman rank* correlation (r_s), a simplified formula for when data are sets of rank-ordered data. Reviewers should *not* demand authors use these alternative computations of the Pearson product-moment correlation, as they are archaic. Essentially, before massive computing power was available in every office, these formulae provided computational shortcuts for people computing correlations by hand or via hand calculator. Researchers using statistical software have no true use for them.

However, there are other correlation coefficients that might be more desirable than r . For example, the *tetrachoric* correlation (Pearson, 1901) is particularly good for examining the relation between raters where the rating is dichotomous (presence or absence of some trait or behavior), while the *polychoric* correlation is designed to examine correlations for ordered-category data (e.g., quality of a lesson plan), another way to measure rater agreement. Importantly, one assumption of these two measures is that while the ratings are dichotomous or categorical, they assume that the latent or underlying variable is continuous (e.g., one can have the presence of particular characteristics of autism to a greater or lesser extent although we might rate them as present/absent). They estimate what the correlation between raters would be if ratings were made on a continuous scale. One final assumption of tetrachoric and polychoric correlations requires that the continuous latent variable follows a bivariate normal distribution.

Robust measures of association, such as *Kendall's tau*, *Spearman's rho*, and *Winsorized* correlations have been developed to attempt to measure association in the presence of violation of assumptions (e.g., presence of outliers), but Wilcox (2008) concluded that none of these measures of association are truly robust in the face of moderate violations of assumptions. The recommended technique currently appears to be the *skipped correlation coefficient* (Wilcox, 2003). This technique, notably more robust than Pearson's r , is a method of measuring the strength of the relation between

two variables while simultaneously ignoring outliers and taking the overall structure of the data into consideration (Wilcox, 2003).

In the case of a truly continuous and dichotomous or polytomous variable (as opposed to a dichotomous or categorical variable that has an underlying continuous distribution), logistic regression may be considered a best practice in determining association between two variables (see Chapter 16, this volume). Although logistic regression has been slow to be adopted in the social sciences, it does carry many benefits, including appropriate mathematical handling of discrete categorical data. One challenge to authors using logistic regression involves the correct interpretation of the measure of association, the *odds ratio* (OR) or index of *relative risk* (RR), the preferred, but more difficult to obtain statistic. A lengthy treatise on these statistics, best practices, and correct interpretation is beyond the scope of this chapter, but see Osborne (2006).

Reviewers should be vigilant about authors who have substantial violations of assumptions and proceed to use these techniques to ameliorate these violations. The best way to deal with violations of assumptions is to deal explicitly with the violation, such as removal of outliers, transformation to improve normality, and so forth. Where this is not possible, robust methods, such as the skipped correlation coefficient, currently are good alternatives. Reviewers seeing authors use archaic correlation coefficients (e.g., point-biserial, phi, or the Spearman rank-order coefficient) should be skeptical as to the authors' quantitative prowess, as this seems to be an indicator of outdated training. Finally, in the specific case of examining inter-rater agreement, alternative nonparametric estimators such as polychoric and tetrachoric correlations are good practices and should be recommended to authors if they are not present).

16. Univariate vs. Multivariate Analyses

Large zero-order correlation tables are often reported in journals. Since the introduction of statistical significance levels⁵ (e.g., $p < .05$), largely attributed to Fisher (1925),⁶ one issue has been maintaining a low rate of Type I errors across an entire set of analyses rather than for each individual analysis. For example, if one reports a correlation table for five variables, the lower triangular matrix contains 10 separate correlations. Assuming researchers are using the traditional $\alpha = .05$ criterion to test the correlations, this means that the Type I error rate across the family of correlations can greatly exceed .05. This seems unacceptable but is routinely done. To combat this issue, early-20th-century statisticians developed corrections for this expanded Type I error rate.

Many different types of corrections for this issue are available (e.g., Bonferroni adjustments), although these types of corrections tend to reduce power and increase the probabilities of a Type II error. Reviewers should expect authors to address inflated Type I error rates in some way, either through a type of correction such as Bonferroni or, preferably, through reducing the number of correlations reported by focusing on the truly important correlations (Curtin & Schulz, 1998) or through the use of multivariate methods. To the extent that $p < .05$ remains an important criterion (and there is a good deal of discussion attempting to reduce reliance on this criterion in the methodology literature, e.g., Killeen, 2008), authors should seek to keep $\alpha = .05$ for each major, substantive hypothesis (e.g., familywise or experimentwise error rate) rather than for each statistical test.

17. Semipartial and Partial Correlations

A *semipartial* correlation is a correlation between two variables (variables 1 and 2), controlling for a third variable (that is correlated in some meaningful way to both of the other variables), with the common notation $sr_{(12,3)}$. Similarly, *partial* correlations are correlations between two variables with the effect of a third removed and use the common notation $pr_{(12,3)}$. However, there is an important

conceptual (and mathematical) difference between the two. For example, consider calculating a correlation between an independent variable (IV) such as motivation and a dependent variable (DV) such as intentions to attend college, controlling for a covariate such as student grades. The semipartial correlation removes the effect of grades from the IV (motivation) but *not* from the DV (intentions). Thus, a semipartial correlation, when squared, communicates the unique variance accounted for in the DV by the IV. This is useful when researchers want to examine which variables make unique contributions to predicting outcomes (e.g., Does motivation predict intentions to attend college once student grades are accounted for?) or which variables account for the most unique variance once others are controlled for (e.g., Does motivation or student grades have a stronger unique effect on intention to attend college?). The semipartial correlation does not, technically, remove the effect of the other variable from the analysis, as it is still in the DV, and authors and researchers should carefully examine interpretations of semipartials so that they are not misinterpreted. Partial correlations, on the other hand, remove the effect of the covariate (e.g., student grades) from both the DV and IV, giving you a more “pure” measure of that relation without the effect of the covariate in either. This is most commonly used when researchers want to remove the effect of confounding or extraneous variables to get a measure of the relation if one could hold other variables constant (e.g., holding grades constant, does motivation have any association with intentions to attend college?).

Finally, when the underlying assumptions of correlation and regression are not met, partialing or controlling for other variables, as mentioned above, can yield unpredictable results. For example, if student grades are not reliably measured and we try to control for them, only a fraction of the variable actually can be removed. The rest of the variance attributable to grades is still in the analysis, and thus if grades and motivation are highly correlated, the effect of motivation may be substantially misestimated. As another example, if the effect of grades is curvilinear and only the linear effect is covaried, then again, part of the effect of grades remains in the analyses, causing misestimation of other effects (curvilinearity is discussed in more detail in Desideratum 22).

Reviewers should pay careful attention that authors use semipartial and partial correlation when appropriate and interpret them accurately. If variables to be used in these types of analyses are not measured reliably, reviewers might insist on discussion of this issue, use of alternative analyses that can correct for low reliability (e.g., structural equation modeling; see Chapter 33, this volume), or reject the analysis as untenable as it is not testing what the authors believe it is testing.

18. Data Cleaning and Data Transformation

Data cleaning is often necessary and needs to be considered when interpreting the analysis results. For example, transformations are tricky to get right computationally (Osborne, 2008a) and also complicate interpretation. First, it is easy to create missing data when doing a transformation (e.g., some scores might be converted to missing data as it is impossible to take the square root or natural log of a negative number). Second, some transformations to reduce skew require the data be *reflected* (the order of the scores is reversed with the highest scores becoming the lowest and the lowest scores becoming the highest) prior to a transformation being applied. If the data are not reflected after the transformation, the correlation computed afterward will be exactly opposite of what it should be (e.g., 0.50 rather than -0.50), which could lead to misinterpretation of the results.

Even assuming a transformation is applied successfully and using best practices, the researcher still is left with one or more transformed variables, an altered version of the original construct. Is it straightforward to say that the correlation with the log of a variable means the same as the correlation with the original variable? Not technically, and probably not in practice, either. So authors need to be clear with readers in interpreting results.

Reviewers need to see evidence that where data transformations are necessary (e.g., highly skewed variable), authors use best practices (see Osborne, 2013 for guidelines on best practices in utilizing some common data transformations) and are mindful that analyses used transformed data when discussing the implications of their results.

19. Interpretation of p Values

For many decades, $p < .05$ has been the primary indicator to a researcher that an effect is “important.” In fact, many authors have been known to place more importance on effects that produce smaller p values. Terms like “more significant” and “highly significant” and different indicators in published research for smaller p values (e.g., using asterisks to indicate significance level: * $p < .05$, ** $p < .01$, *** $p < .001$). There are two issues that reviewers need to attend to: misuse of p as a proxy for importance of an effect and misinterpretation of p .

It might not be surprising that some researchers use p as a proxy for importance of an effect or for effect size itself (discussed in more detail in Desideratum 20). One of the primary factors in the magnitude of p is effect size. However, other factors also help determine p ; in fact, sample size is a primary determiner of p . Therefore, p is not usable for this purpose. Reviewers need to ensure that authors disentangle effect size/importance with significance level.

Reporting confidence intervals is a best practice and aids in the interpretations of p values. Because the distribution of r is skewed, the confidence interval around r is not symmetrical. To compute a confidence interval around r , one must use a Fisher’s r -to- z transformation. Confidence intervals not only aide in interpretation and help with the issue of disentangling importance/effect size with significance level but also aid in comparisons of reported correlations. As such, reviewers need to ensure that authors report confidence intervals whenever they report and interpret correlations.

Once statistically significant effects are identified, effect sizes and confidence intervals around effect sizes should guide discussion as to importance of the findings (see Chapter 6, this volume). Authors should focus on effect sizes in interpreting results, and reviewers should closely monitor manuscripts for author hyperbole.

Finally, several authors have proposed modern alternatives to p , given that p usually does not provide the information researchers seek. One of the most promising alternatives is Killeen’s (2008) $p_{(rep)}$, the *probability of replication*, a statistic that is directly related to what most researchers are interested in, and easily calculated from information most statistical software provides.

20. Effect Size Measures

Effect sizes are simple to calculate in the context of correlation and regression analyses. Unlike other types of analyses such as ANOVA and t tests, correlation coefficients are themselves effect sizes, as are β s and multiple R s in regression (for more information on this point, see, e.g., Cohen, 1988; Cohen et al., 2003). Effect sizes tend to come in two general categories: a standardized index of strength and a percentage of variance accounted for. R , β , sr , pr , and r are all indices of strength (as are analogous effect size indexes for ANOVA type designs, such as d , f , and ω^2). Because indices of strength tend to be on different scales, it is good practice for authors to report effect sizes that represent the proportion of variance accounted for. In the case of association, r^2 , sr^2 , pr^2 , and R^2 are appropriate. Some authors (e.g., Thompson, 2002) have argued that effect sizes should be accompanied by confidence intervals for effect sizes. Because this last technique is not commonly available via statistical packages at this time (it is available through the R statistical package, freely available at www.r-project.org), its use should be seen as a suggestion rather than a requirement.

Reviewers should “reality check” authors’ claims about strength of association. Authors might over-exaggerate the strength of the effects in their research in hopes of making the projects more publishable. For example, correlations of $r = .30$ only represent 9% variance accounted for, and as such, do not generally qualify as “strong” effects or “good” support for consistency. Rules of thumb for what constitutes a strong, moderate, or weak effect for various effect size indices are widely disseminated (e.g., Cohen, 1962, 1988) and are also available on the internet.

21. Curvilinear Relations and Interactions

Few researchers in the social sciences examine or discuss curvilinear effects. Is this because there are so few curvilinear effects or because researchers fail to look for them? Until recently, it was difficult to test for curvilinear effects explicitly within statistical software frameworks. However, most popular modern statistical packages include simple diagnostic options to identify possible curvilinearity (e.g., residual plots) and contain curvilinear regression options, both of which allow researchers to test for curvilinear effects. When found, these effects are usually best conveyed to readers as graphs that show the nature of the curvilinearity; thus, reviewers should expect a graph to accompany the effect description/interpretation, and best practices in graphing should be followed (see Osborne, 2017, for detailed examples of modeling curvilinearity).

22. Causal Inferences

The “cardinal sin” of correlational analysis is to make causal statements or inferences when the study design does not warrant it. For example, one might read, “Because availability of instructional technology is associated with stronger growth in student achievement, schools should make stronger efforts to include instructional technology in the daily life of students.” Anyone having taken a basic research methods course should remember that any relation could have at least three possible causal bases: (a) variable A could cause B, (b) variable B could cause A, and (c) the relation between A and B could be caused by a third variable, C, that is not measured. Authors should be careful to respect the limits of correlational methodologies and exercise appropriate caution regarding causal inferences or statements. That being said, correlational design studies that can be construed as quasi-experimental, such as propensity score matching (see Chapter 28 in this volume), can yield causal inferences under the right assumptions.⁷

Notes

- 1 Exploratory factor analysis (see Chapter 8, this volume) is often difficult to defend because of low replicability and lack of inferential statistics.
- 2 Stratified sampling refers to the process of grouping members of the population into relatively homogeneous subgroups before sampling. The strata should be mutually exclusive: every element in the population must be assigned to only one stratum, and ideally, every element of the population should be represented by a stratum.
- 3 Many authors have acknowledged the significant issues with null hypothesis statistical testing (NHST) and some (see particularly Killeen, 2008) have proposed alternatives such as the probability of replication as a more interesting/useful replacement. Interested readers should examine Thompson (1993) and Fidler (2005) as an introduction to the issues.
- 4 Design effects are often calculated and disseminated in large, standardized surveys with complex sampling designs, such as government surveys or samples. They are methodologically challenging to calculate, but if known, can be incorporated into analyses as an adjustment to effect sizes and significance tests so as to account for the artificially small standard errors present in certain types of samples, such as those with nested data. Multilevel modeling is generally a more elegant and effective way to deal with this issue.
- 5 Those interested in the history leading to the origins of $p < .05$ should read Cowles and Davis’s (1982) excellent history of significance testing.

- 6 Although the concept of formal, norm-guided significance testing itself is traceable to Student (1908) and Wood and Stratton (1910).
- 7 Because this chapter is focusing on relatively simple associational analyses, we note that simple associational analyses should not be couched in causal terms. However, more sophisticated analytic techniques, such as structural equation models based on strong theory can use correlational data to support or refute hypotheses. However, at some point causal statements need to be assessed via methodologies (e.g., double blind experimental studies) that are designed to test causal inference more effectively.

Acknowledgments

The authors would like to acknowledge and thank Sr. Kathleen Cash, Ph.D. for her feedback on this chapter.

References

- Aiken, L. S., & West, S. G. (2010). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145–153.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. New York: Routledge.
- Cole, J. C. (2008). How to deal with missing data. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 214–238). Thousand Oaks, CA: Sage.
- Cowles, M. D., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, 37, 553–558.
- Curtin, F., & Schulz, P. (1998). Multiple correlations and Bonferroni's correction. *Biological Psychiatry*, 44, 775–777.
- Deemer, W. L. (1947). The power of the *t* test and the estimation of required sample size. *Journal of Educational Psychology*, 38, 329–342.
- Fidler, F. (2005). From statistical significance to effect estimation: Statistical reform in psychology, medicine, and ecology. Doctoral Dissertation, University of Melbourne.
- Fidler, F., & Cumming, G. (2008). The new stats: Attitudes for the 21st century. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 1–14). Thousand Oaks, CA: Sage.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.
- Killeen, P. R. (2008). Replication statistics. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 103–124). Thousand Oaks CA: Sage.
- Lee, E. S., Forthofer, R. N., & Lorimor, R. J. (1989). *Analyzing complex survey data*. Newbury Park, CA: Sage.
- McCoach, D. B., & Adelson, J. L. (2010). Dealing with dependence (Part I): Understanding the effects of clustered data. *Gifted Child Quarterly*, 54, 152–155.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw Hill.
- Osborne, J. W. (2003). Effect sizes and the disattenuation of correlation and regression coefficients: Lessons from educational psychology. *Practical Assessment, Research, and Evaluation*, 8(99).
- Osborne, J. W. (2006). Bringing balance and technical accuracy to reporting odds ratios and the results of logistic regression analyses. *Practical Assessment Research & Evaluation*, 11(7).
- Osborne, J. W. (2008a). Best practices in data transformation: The overlooked effect of minimum values. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 197–204). Thousand Oaks, CA: Sage.
- Osborne, J. W. (2008b). Sweating the small stuff in educational psychology: How effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educational Psychology*, 28, 1–10.
- Osborne, J. W. (2008c). A brief introduction to hierarchical linear modeling. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 445–450). Thousand Oaks, CA: Sage.
- Osborne, J. W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Thousand Oaks, CA: Sage Publishing. ISBN: 9781412988018.
- Osborne, J. W. (2017). *Regression and linear modeling: Best practices and modern methods*. Thousand Oaks, CA: Sage.
- Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, Research, and Evaluation*, 9(6).
- Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research, and Evaluation*, 8(2).
- Pearson, K. (1901). Mathematical contribution to the theory of evolution. VII: On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London*, A, 195, 1–47.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction*. Fort Worth, TX: Wadsworth/Thomson Learning.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (vol. 1). Thousand Oaks, CA: Sage.
- Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Counseling and Clinical Psychology*, 58, 646–656.
- Student. (1908). The probable error of a mean. *Biometrika*, 6, 1–25.

- Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (6th ed.). Harlow: Pearson.
- Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, 61, 361–377.
- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, 31, 24–31.
- Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, 6, 100–116.
- Wilcox, R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.
- Wilcox, R. (2008). Robust methods for detecting and describing associations. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 263–279) Thousand Oaks, CA: Sage.
- Wood, T. B., & Stratton, F. J. M. (1910). The interpretation of experimental results. *Journal of Agricultural Science*, 3, 417–440.
- Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education*, 67, 55–68.

6

Effect Sizes and Confidence Intervals

Fiona Fidler and Geoff Cumming

An *effect size* is simply an amount of something of interest. It can be as simple as a mean, a percentage increase, or a correlation; or it may be a regression weight, a standardized measure of a difference, or the percentage of variance accounted for. Most research questions in the social sciences are best answered by finding estimated effect sizes, meaning *point estimates* of the true effect sizes in the population. Grissom and Kim (2012) provided a comprehensive discussion of effect sizes, and ways to calculate effect size estimates. Our discussion mainly focuses on experimental designs, but much of the discussion is relevant also to other types of research.

A *confidence interval* (CI), most commonly a 95% CI, is an *interval estimate* of a population effect size, which indicates the *precision* of the point estimate. The *margin of error* (MoE) is the length of one arm of a CI. The most common CIs are symmetric, and for these the MoE is half the total length of the CI. The MoE is our measure of precision. Cumming and Finch (2005) provided an introduction to CIs and their advantages.

In the social sciences, statistical analysis is still dominated by null hypothesis significance testing (NHST). However, there is extensive evidence that NHST is poorly understood, frequently misused, and often leads to incorrect conclusions. It is important and urgent that social scientists shift from relying mainly on NHST to using better techniques, including especially effect sizes, CIs, and meta-analysis. The changes needed are discussed in detail by Cumming (2012, 2014) and Kline (2013).

The most important current development in research methodology is the rise of Open Science, a set of practices designed to increase the openness, integrity, and reproducibility of research. One precursor was the classic article by Ioannidis (2005) that identified three problems which, together, suggested that “Most published research results are false”—to quote from the article’s title. The problems are (1) selective publication, especially of results that are statistically significant; (2) the imperative to achieve $p < .05$, which prompts researchers to select and tweak what they report until they can claim statistical significance; and (3) the widespread belief that once a result has achieved $p < .05$ and been published it need not be replicated. Ioannidis identified over-reliance on NHST as the common factor underlying all three problems. Another key contribution was that of Simmons, Nelson, and Simonsohn (2011), who explained that the very large number of choices typically made by researchers as they analyzed their data and chose what to report meant that statistical significance could usually be found, no matter what the results. Expanding on the second problem of Ioannidis, they described many of those choices as *questionable research practices*, including for

example, testing some extra participants after seeing the data, dropping some aberrant data points, and selecting which measures or comparisons to highlight. The use of questionable research practices to achieve $p < .05$ is called *p-hacking*, which is a pernicious and totally unacceptable practice. Widespread disquiet about *p* values and how they are used led the American Statistical Association to make a strong statement about their shortcomings (Wasserstein & Lazar, 2016).

The Center for Open Science (COS, cos.io) was founded in 2013 to promote Open Science, including especially the conduct of replications. It provides the Open Science Framework (OSF, osf.io), which is an invaluable and freely available online resource for researchers wishing to adopt Open Science practices. One fundamental Open Science practice is *preregistration* of a research plan, including a data analysis plan, in advance of running a study. Preregistering, then following the plan closely, helps ensure that planned and exploratory analysis can be clearly distinguished. It should also lead to reporting the study, whether in a journal or in an enduring accessible repository such as OSF, whatever results the study obtains. This would be a large step towards overcoming Ioannidis's first problem of selective publication.

Adopting estimation and meta-analytic perspectives helps us overcome the three Ioannidis problems, but more is required—questionable research practices such as the dropping of apparently aberrant results, or reporting only some selected comparisons, can be just as damaging when using CIs as with NHST. The full range of Open Science practices is needed.

Before assessing a manuscript, a reviewer should be familiar with the target journal's policies on Open Science issues, in particular the *Transparency and Openness Promotion (TOP) Guidelines*, which are available at cos.io/top/. In relation to various Open Science practices that we sketch in Table 6.1,

Table 6.1 Desiderata for Effect Sizes and Confidence Intervals.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. The main questions to be addressed are formulated in terms of estimation and not simply null hypothesis significance testing.	I
2. Previous research literature is discussed in terms of effect sizes, confidence intervals, and from a meta-analytic perspective.	I
3. The rationale for the design—whether experimental or otherwise—and procedure is explained and justified in terms of appropriateness for obtaining precise estimates of the target effect sizes.	I, M
4. The dependent variables are described and operationalized with the aim that they should lead to good estimates of the target effect sizes.	M
5. Where possible a detailed research plan including data analysis plan was preregistered, then followed.	M
6. Results are presented and analyzed in terms of point estimates of the effect sizes.	R
7. The precision of effect size estimates is presented and analyzed in terms of confidence intervals.	R
8. Wherever possible, results are presented in figures, with confidence intervals.	R, D
9. Effect sizes are given substantive interpretation.	D
10. Confidence intervals are given substantive interpretation.	D
11. Meta-analytic thinking is used to interpret and discuss the findings. Replication is considered.	D
12. Where possible, full details of the materials and procedure are made openly available online.	D
13. Where possible, the data are made openly available online.	D

* I = Introduction, M = Method, R = Results, D = Discussion.

such as preregistration, or provision of open data, a journal might encourage the practice, or even require it. It may offer badges, created by the Center for Open Science (tiny.cc/badges), to recognize articles that provide open data or open materials, or that preregistered the study being reported. Any manuscript needs to be assessed against such policies of the target journal.

Our aim in this chapter is to assist authors and manuscript reviewers to make the vital transition from over-reliance on NHST to more informative statistical methods and Open Science.

1. Formulation of Main Questions as Estimation

An astronomer wishes to know the age of the Earth; a chemist measures the boiling point of an interesting new substance: These are the typical questions of science. Correspondingly, in the social sciences we wish to estimate how seriously divorce disrupts adolescent development, or the effect of a type of psychotherapy on depression in the elderly. The chemist reports her result as, for example, $27.35 \pm 0.02^\circ\text{C}$, which signals that 27.35 is the point estimate of the boiling point, and 0.02 is the precision of that estimate. Correspondingly, it is most informative if the psychologist reports the effect of the psychotherapy as an effect size—the best estimate of the amount of change the therapy brings about—and a 95% CI to indicate the precision of that estimate. This approach can be contrasted with the impoverished dichotomous thinking (there is, or is not, an effect) that is prompted by NHST.

In expressing their aims, authors should use language such as:

- We estimate the extent of . . .
- Our aim is to find how large an effect . . . has on . . .
- We investigate the nature of the relationship between . . . and . . .
- We will estimate how well our model fits these data . . .

Expressions like these naturally lead to answers that are effect size estimates. Contrast these with statements like, “We investigated whether this treatment has an effect,” which suggests that a mere dichotomous yes-or-no answer would suffice. Almost certainly the new treatment has *some* effect; our real concern is whether that effect is tiny, or even negative, or is positive and usefully large. It is an estimate of effect size that answers these questions.

Examine the wording used to express the aims and main questions of the manuscript, especially in the abstract and introduction, but also in the title. Replace any words that betray dichotomous thinking with words that ask for a quantitative answer.

2. Previous Literature

Traditionally, reviews of past research in the social sciences have focused on whether previously published studies have, or have not, found a statistically significant effect. That is an impoverished and misleading approach, which ignores the sizes of effects observed, and the fact that many negative results are likely to have been Type II errors attributable to low statistical power.

Past research should, wherever possible, be discussed in terms of the point and interval estimates obtained for the effects of interest. Most simply, an effect size is a mean or other measurement in the original measurement units: The average extent of masked priming was 27 ms; the mean improvement after therapy was 8.5 points on the Beck Depression Inventory; the regression of annual income against time spent in education was 3,700 dollars per year. Alternatively, an effect size measure may be units-free: After therapy, 48% of patients no longer met the criteria for the initial clinical diagnosis; the correlation between hours of study and final grade was .52; the odds ratio for risk of unemployment in young adults not in college is 1.4, for males compared with females. Some effect

size measures indicate percentage of variance accounted for, such as R^2 , as often reported in multiple regression, and η^2 or ω^2 , as often reported with ANOVA. An important class of effect size measures are *standardized effect sizes*, including Cohen's d and Hedges's g . These are differences—typically between an experimental and a control group—measured in units of some relevant standard deviation (SD), for example the pooled SD of the two groups. Cumming and Finch (2001) explained Cohen's d and how to calculate CIs for d . The most appropriate effect size measure needs to be chosen for each research question, in the context of the research design. Grissom and Kim (2012) is an excellent source of assistance with the choice, calculation and presentation of a wide variety of effect size measures.

The introduction to the manuscript should focus on the effect size estimates reported in past research, to provide a setting for the results to be reported. It is often helpful to combine the past estimates, and *meta-analysis* allows that to be done quantitatively. Hunt (1997) gave a general introduction to meta-analysis, and an explanation of its importance. Borenstein, Hedges, Higgins, and Rothstein (2009), Cooper (2010), and Cumming (2012, ch. 7–9) provided guidance for conducting a meta-analysis, and Chapter 19 of this volume discusses meta-analysis in more detail.

Figure 6.1 is a *forest plot*, which presents the results of 12 studies, and their combination by meta-analysis. The result of each study is shown as a point estimate, with its CI. The result of the meta-analysis is a weighted combination of the separate point estimates, also shown with its CI. This CI on the result is usually much shorter, indicating greater precision, as we would expect given that results are being combined over multiple studies. Some medical journals now routinely require the introduction to each empirical article to cite a meta-analysis—or, if none is available, wherever possible to carry out and report a new meta-analysis—as part of the justification for undertaking new research. That is a commendable requirement. Forest plots summarize a body of research in a compact and clear way; they are becoming common in medicine, and should be used more widely.

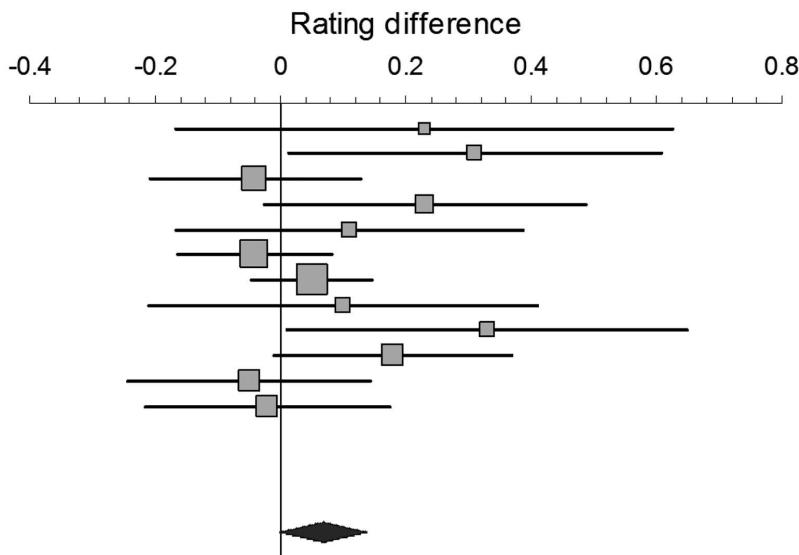


Figure 6.1 A Forest Plot, Showing the Results of 12 Studies that Estimated the Mean Rating Difference between Two Types of Stimuli.

Squares indicate point estimates, and error bars the 95% CIs. The sizes of the squares indicate approximately the weights of the different studies in the meta-analysis. The result of the meta-analysis is displayed as a diamond, whose horizontal extent is the 95% CI.

3. Experimental Design and the Precision of Estimates

Traditionally, statistical power estimates have been used to guide selection of the sample size N required if a planned study is to have a reasonable chance of identifying an effect of a specified size, should this exist. The power approach was advocated by Jacob Cohen, and his book (Cohen, 1988) provided tables and advice (see also Chapter 27, this volume). An Internet search readily identifies freely available software to carry out power calculations, including G*Power (tiny.cc/gpower3). The power approach can be useful, but statistical power is defined in the context of NHST, and has meaning only in relation to a specified null hypothesis. Null hypotheses are almost always statements of zero effect, zero difference, or zero change. Rarely is such a null hypothesis at all plausible, and so it a great advantage of CIs that no null hypothesis need be formulated. In addition, CIs offer an improved approach to selecting N .

An important advance in statistical practice is routine use of precision, meaning the MoE, in planning a study, as well as in discussion and interpretation of results. Cumming (2012, ch. 13) described such a *precision for planning* approach that avoids NHST and the need to choose a null hypothesis. It is based on calculation of what sample size is needed to give a CI with a chosen target length: How large must N be for the expected 95% CI to be no longer than, for example, 60 ms? Given a chosen experimental design, what sample size is needed for the expected MoE to be 0.2 units of Cohen's d ? For two independent groups each of size N , Figure 6.2 shows a graph of required N against expected MoE, expressed in units of σ , the population SD.

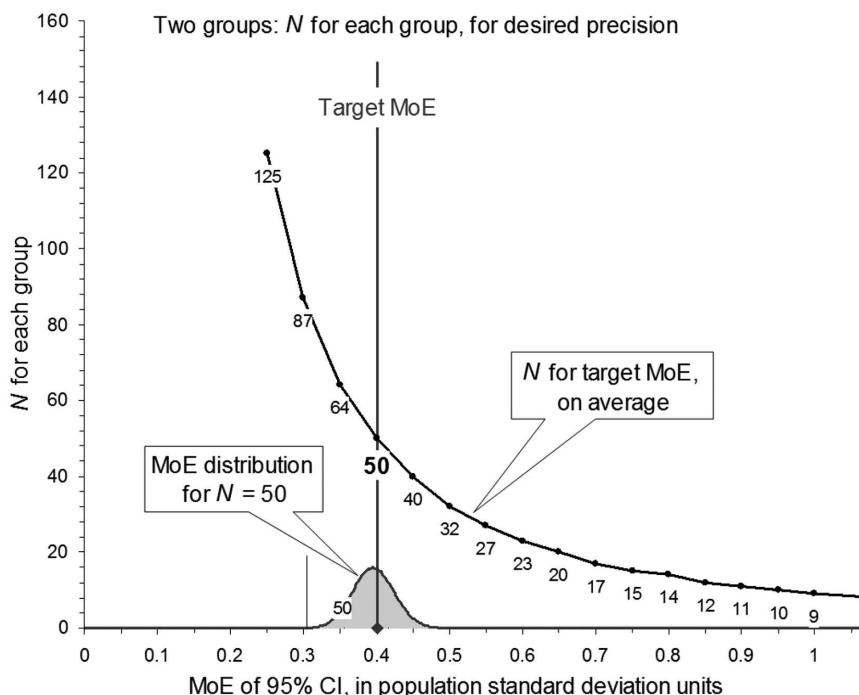


Figure 6.2 Graph Required for Precision for Planning.

The curve indicates the N required by a two independent groups study, each group of size N , for MoE of the 95% CI on the difference between the group means to be as shown on the horizontal axis, on average. The vertical cursor marks MoE = 0.4, in units of σ , the population SD. The shaded curve is the distribution of MoE values over a very large set of studies all having $N = 50$.

Justification of the experimental design and chosen sample size should appear as part of the rationale at the end of the Introduction section, or in the Method section. It is often omitted from journal articles, having been overlooked by authors and reviewers, or squeezed out by strict word limits. Providing such justification is, however, especially important in cases where using too small a sample is likely to give estimates so imprecise that the research is scarcely worth doing, and may give misleading results. It is ethically problematic to carry out studies likely to give such inaccurate results. The converse—studies with such a large sample of participants that effects are estimated with greater precision than is necessary—tend to be less common, but may be ethically problematic if they subject a needlessly large number of participants to an uncomfortable or time-consuming procedure. The best way to justify a proposed design and sample size is in terms of the precision of estimates—the expected MoE—likely to be given by the results.

4. Dependent Variables

Specifying the experimental questions in terms of estimation of effect sizes leads naturally to choice of the dependent variables (DVs), or measures, that are most appropriate for those questions. Choose the operationalization of each DV that is most appropriate for expressing the effect sizes to be estimated, and that has adequate measurement properties, including reliability and validity. The aim is to choose measures that (1) relate substantively most closely to the experimental questions, and therefore will give results that are meaningful and interpretable; and (2) are most likely to give precise estimates of the targeted population effects.

In the Introduction section there may be discussion of methods used in past research, and this may help guide the choice of measures. In the Methods section there may be reference to published articles that provide information about the development of particular measures, and their psychometric properties. One important consideration is that the results to be reported should be as comparable as possible with previous research, and likely future research, so that meta-analytic combination over studies is as easy as possible. It can of course be a notable contribution to develop and validate an improved measure, but other things being equal it is advantageous to use measures already established in a field of research.

Choice of measures is partly a technical issue, with guidance provided by psychometric evidence of reliability and validity in the context of the planned experiment. It is also, and most importantly, a substantive issue that requires expert judgment by the researchers: The measures must tap the target concepts, and must give estimates of effects that can be given substantive and useful interpretation in the research context.

5. Preregistration of a Research Plan

The OSF makes it easy for researchers or students to preregister a detailed research plan, including a data analysis plan, in an online repository where it is date-stamped and cannot be changed. It can be kept confidential, or at any time the researcher can make it open to all. Preregistration has long been required for drug trials, but only recently has it become recognized as important in the social sciences. Wagenmakers et al. (2012) explained the benefits and importance of preregistration. Authors are free to submit a study that had been preregistered to any journal, but, since 2014, *Psychological Science* has offered a badge for preregistered studies. Increasing numbers of journals are encouraging preregistration and offering the badge.

If a manuscript claims preregistration, a link must be provided to the plan that was lodged before data collection commenced, and the study must have been conducted and analyzed in accordance with that plan. Data exploration beyond the planned analysis may be acceptable, but any results

found by exploration may easily be cherry picked, mere capitalization on chance, and are at best speculations, perhaps for further investigation.

6. Results: Effect Sizes

The main role of the Results section is to report the estimated effect sizes that are the primary outcomes of the research. We mentioned in Desideratum 2 the wide range of possible effect size measures, and emphasized that many of these are as simple and familiar as means, percentages and correlations. In many cases it is possible to transform one effect size measure into a number of others; Kirk (1996, 2003) provided formulas for this purpose. A correlation, for example, can be transformed into a value of Cohen's d . It is a routine part of meta-analysis to have to transform effect size estimates reported in a variety of ways into some common measure, as the basis for conducting the meta-analysis. In medicine, odds ratio or log odds ratio are frequently used as the common effect size measure, but in social science Cohen's d , or Pearson's r correlations are frequently chosen as the basis for meta-analysis.

Often it may be useful to present results in the original measurement scale of a DV, for simplicity and ease of interpretation, and also in some standardized form to assist comparison of results over different studies, and the conduct of future meta-analysis. For example, an improvement in depression scores might be reported as mean change in score on the Beck Depression Inventory (BDI), because such scores are well known and easily interpreted by researchers and practitioners in the field. However if the improvement is also reported as a Cohen's d value the result is easily compared with, or combined with, the results of other studies of therapy, even where they have used other measures of depression. Similarly, a regression coefficient could be reported both in raw form, to assist understanding and interpretation, and as a standardized value, to assist comparison across different measures and different studies. In any case it is vital to report SDs, and mean square error values, so that later meta-analysts have sufficient information to calculate whichever standardized effect size measures they require.

A standardized measure of difference, such as Cohen's d , can be considered simply as a number of standard deviations. It is in effect a z score. It is important to consider which SD is most appropriate to use as the basis for standardization. Scores on the BDI, and changes in BDI scores, could be standardized against a published SD for the BDI. The SD unit would then be the extent of variation in some BDI reference population. That SD would have the advantage of being a stable and widely available value. Similarly, many IQ measures are already standardized to have a SD of 15. Alternatively, a change in BDI score could be expressed in units of the pre-test SD in our sample of participants. That would be a unit idiosyncratic to a specific study, and containing sampling error, but it might be chosen because it applies to the particular patient population we are studying, rather than the BDI reference population. As so often is the case in research, informed judgment is needed to guide the choice of SD for standardization. When a manuscript reports a Cohen's d value, or any other standardized measure, it is essential that it make clear what basis was chosen for standardization.

It may be objected that much research has the aim not of estimating how large an effect some intervention has, but of testing a theory. However, theory testing is most informative if considered as a question of estimating goodness of fit, rather than of rejecting or not rejecting a hypothesis derived from the theory. A goodness of fit index, which may be a percentage of variance, or some other measure of distance between theoretical predictions and data, is an effect size measure, and point and interval estimates of goodness of fit provide the best basis for evaluating how well the theory accounts for the data (Velicer et al., 2008).

7. Results: Confidence Intervals

Following the *Publication Manual* of the American Psychological Association (APA, 2010) we recommend the following style for reporting CIs in text. At the first occurrence in a paragraph write: “The mean decrease was 34.5 ms, 95% CI [12.0, 57.0], and so . . .” On later occasions in the paragraph, if the meaning is clear write simply: “The mean was 4.7 cm [−0.8, 10.2], which implies that . . .” or “The means were 84% [73, 92] and 65% [53, 76], respectively . . .,” or “The correlation was .41 [.16, .61] . . .” The units should not be repeated inside the square brackets. Note that in the last example, which gives the 95% CI on Pearson’s $r = .41$, for $N = 50$, the interval is not symmetric about the point estimate; asymmetric intervals are the norm when the variable has a restricted range, as in the cases of correlations and proportions.

We recommend general use of 95% CIs, for consistency and to assist interpretation by readers, but particular traditions or special circumstances may justify choice of 99%, 90%, or some other CIs. If an author elects to use CIs with a different level of confidence, then that should be stated in every case: “The mean improvement was 1.20 scale points, 90% CI [−0.40, 2.80].”

In a table, 95% CIs may similarly be reported as two values in square brackets immediately following the point estimate. Alternatively, the lower and upper limits of the CIs may be shown in separate labeled columns.

Cumming and Finch (2001) and Cumming (2012, ch. 11) explained how to calculate CIs for Cohen’s d . Grissom and Kim (2012) provided advice on how to calculate CIs for many measures of effect size. Calculation of CIs can be straightforward, or may best be accomplished using computer-intensive methods, such as bootstrapping. Helpful software is becoming increasingly available (Cumming, 2014, p. 25).

8. Figures with Confidence Interval Error Bars

Whenever possible, researchers should provide figures that include 95% CIs. Cumming and Finch (2005) discussed the presentation and interpretation of error bars in figures. A serious problem is that the familiar graphic used to display error bars in a figure, as shown in Figure 6.3, can have a number of meanings. The bars could indicate SD, standard error (SE), a 95% CI, a CI with some other level of confidence, or even some other measure of variability. Cumming, Fidler, and Vaux (2007) described and discussed several of these possibilities. The most basic requirement is that any figure with error bars must include a clear statement of what the error bars represent. A reader can make no sense of error bars without being fully confident of what they show, for example 95% CIs, rather than SDs or SEs.

CIs are interval estimates and thus provide inferential information about the effect size of interest. CIs are therefore almost always the intervals of choice. In medicine it is CIs that are recommended and routinely reported. In some research fields, however, including behavioral neuroscience, SE bars (error bars that extend one SE below and one SE above a mean) are often shown in figures. When sample size is at least about 10, SE bars are about half the length of the 95% CI, so it is easy to translate visually between the two. But SE bars are not accurately and directly inferential intervals, so CIs should almost always be preferred.

Figure 6.3 shows means with CIs for a hypothetical two-group experiment with a repeated measure. A treatment group was compared with a control group, and three applications of an anxiety scale provided pre-test, post-test, and follow-up measures. The figure illustrates several important issues. First, a knowledgeable practitioner might feel that the CIs are surprisingly and discouragingly long, despite the reasonable group sizes ($N = 23$ and 26). It is an unfortunate reality across the social sciences that error variation is usually large. CI length represents accurately the uncertainty

inherent in a set of data, and we should not shoot the messenger by being critical of CIs themselves for being too long. The problem is NHST, with its simplistic reject or don't reject outcome, which may delude us into a false sense of certainty, when in fact much uncertainty remains. Cohen (1994, p. 1002) said, "I suspect that the main reason they [CIs] are not reported is that they are so embarrassingly large!" We should respond to the message of large error variation by making every effort to improve experimental design and use larger samples, but must acknowledge the true extent of uncertainty by reporting CIs wherever possible.

Cumming and Finch (2005) provided rules of eye to assist interpretation of figures such as Figure 6.3. For means of two independent groups, the extent of overlap of the two 95% CIs gives a quick visual indication of the approximate p value for a comparison of the means. If the intervals overlap by no more than about half the average of the two MoEs, then $p < .05$. If the intervals have zero overlap—the intervals touch end-to-end—or there is a gap between the intervals, then $p < .01$. In Figure 6.3 the control and treatment means at pre-test, for example, overlap extensively, and so p is considerably greater than .05. At post-test, however, the intervals have only a tiny overlap, so at this time point p for the treatment vs. control comparison is approximately .01. At follow-up, overlap is about half the length of the average of the two overlapping arms (the two MoEs), and so p is approximately .05.

It is legitimate to consider overlap when the CIs are on independent means, but when two means are paired or matched, or represent a repeated measure, overlap of intervals is *irrelevant* to the comparison of means, and may be misleading. Further information is required, namely the correlation between the two measures, or the SD of the *differences*. For this reason it is not possible to assess in Figure 6.3 the p value for any within-group comparison, such as the pre-test to

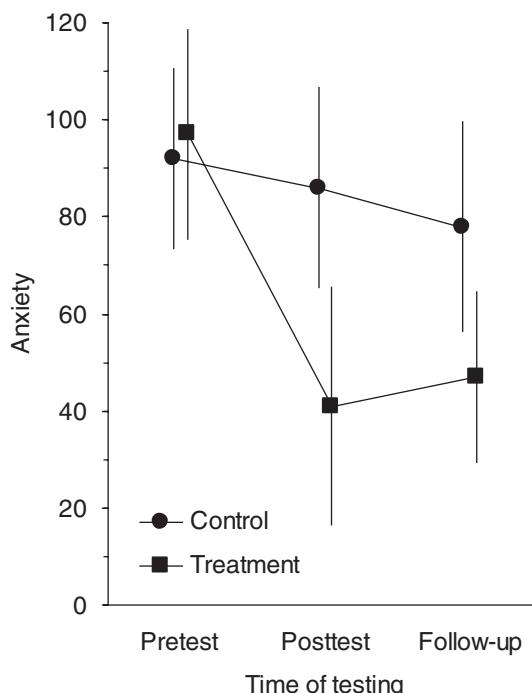


Figure 6.3 Mean Anxiety Scores and 95% Confidence Intervals (CIs) for a Fictitious Study.

The scores compare a Treatment ($N = 23$) and a Control ($N = 26$) group, at each of three time points: pre-test, post-test, and follow-up. Means have been displaced slightly so all CIs can be clearly seen.

post-test change for the treatment group. Belia, Fidler, Williams, and Cumming (2005) reported evidence that few researchers appreciate the importance of the distinction between independent and dependent means when interpreting error bars. If CIs in figures are to be used to inform the interpretation of data—as we advocate—it is vital that figures make very clear the status of each independent variable. For between-subject variation, or independent means, intervals can be directly compared. For within-subject variation, a repeated measure, or dependent means, intervals may not be compared.

It is a problem that many current software packages do not sufficiently support the preparation of figures with error bars. In Figure 6.3, for example, the means are slightly offset horizontally so that all CIs can be seen clearly, but few packages make it easy to do this. One solution is to use Microsoft Excel. Figure 6.3 was prepared as an Excel scatterplot, which requires the horizontal and vertical coordinates for each point to be specified, so means can readily be displayed with a small horizontal offset.

In summary, the Results section should report point and interval estimates for the effect sizes of interest. Figures, with 95% CIs shown as error bars, should be presented wherever that would be informative. Every figure must make clear what error bars represent, and must describe the experimental design so a reader can understand whether each independent variable varies between or within subjects.

9. Interpretation of Effect Sizes

A primary purpose of the Discussion section is to present a substantive interpretation of the main effect size estimates, and to draw out the implications. One unfortunate aspect of NHST is that the term *significant* is used with a technical meaning—a small p value was obtained—whereas in common language the word means “important.” Kline (2013) recommended the word simply be dropped, so that if a null hypothesis is rejected we would say “a statistical difference was obtained.” The common practice of saying “a significant difference was obtained” almost insists that a reader regard the difference as important, whereas it may easily be small and of trivial importance, despite yielding a small p value. Judging whether an effect size is large or important is a key aspect of substantive interpretation, and requires specialist knowledge of the measure and the research context. We recommend that, if reporting NHST, either avoid the term “significant,” as Kline recommends, or make its technical meaning clear by saying “statistically significant.” When discussing the importance of a result, use words other than “significant,” perhaps including “notable,” “clinically important,” or “educationally important.”

Cohen (1988, pp. 12–14) suggested reference values for the interpretation of some effect size measures. For example, for Pearson correlation he suggested that values of .1, .3, and .5 can be regarded as small, medium, and large, respectively; and for Cohen’s d he suggested similar use of .2, .5, and .8. However, he stated that his reference values were arbitrary, and “were set forth throughout with much diffidence, qualifications, and invitations not to employ them if possible” (Cohen, 1988, p. 532). Sometimes numerically tiny differences may have enormous theoretical importance, or indicate a life-saving treatment of great practical value. Conversely a numerically large effect may be unsurprising and of little interest or use. Knowledgeable judgment is needed to interpret effect sizes (How large? How important?), and a Discussion section should give reasons to support the interpretations offered, and sufficient contextual information for a reader to come to an independent judgment.

10. Interpretation of Confidence Intervals

A manuscript should not only report CIs, but also use them to inform the interpretation and discussion of results. The correct way to understand the level of confidence, usually 95%, is in relation

to indefinitely many replications of an experiment, all identical except that a new sample is taken each time. If the 95% CI is calculated for each experiment, in the long run 95% of these intervals will include the population mean μ , or other parameter being estimated. For our sample, or any particular sample, the interval either does or does not include μ , so the probability that this particular interval includes μ is 0 or 1, although we will never know which. It is misleading to speak of a probability of .95, because that suggests the population parameter is a variable, whereas it is actually a fixed but unknown value.

Here follow some ways to think about and interpret a 95% CI (see also Cumming, 2012; Cumming & Finch, 2005).

- The interval is one from an infinite set of intervals, 95% of which include μ . If an interval does not contain μ , it probably only just misses.
- The interval is a set of values that are *plausible* for μ . Values outside the interval are relatively implausible—but not impossible—for μ . (This interpretation may be the most practically useful.)
- We can be 95% confident that our interval contains μ . If in a lifetime of research you calculate numerous 95% CIs in a wide variety of situations, overall, around 95% of these intervals will include the parameters they estimate, and 5% will miss.
- Values around the center of the interval are the best bets for μ , values towards the ends (the lower and upper limits) are less good bets, and values just outside the interval are even less good bets for μ (Cumming, 2007).
- The lower limit is a likely lower bound of values for μ , and the upper limit a likely upper bound.
- If the experiment is replicated, there is on average about an 83% chance that the sample mean (the point estimate) from the replication experiment will fall within the 95% CI from the first experiment (Cumming, Williams, & Fidler, 2004). In other words, a 95% CI is approximately an 83% *prediction interval* for the next sample mean.
- The MoE is a measure of precision of the point estimate, and is the likely largest error of estimation, although larger errors are possible.
- If a null hypothesized value lies outside the interval, it can be rejected with a two-tailed test at the .05 level. If it lies within the interval, the corresponding null hypothesis cannot be rejected at the .05 level. The further outside the interval the null hypothesized value lies, the lower is the p value (Cumming, 2007).

The last interpretation describes the link between CIs and NHST: Given a CI it is easy to note whether any null hypothesized value of interest would be rejected, given the data. Note, however, the number and variety of interpretations of a CI that make no reference to NHST. We hope these will become the predominant ways researchers think of CIs, as CIs replace NHST in many situations.

Authors may choose any of the options above to guide their use of CIs to interpret their results. As the *Publication Manual* recommends, “wherever possible, base discussion and interpretation of results on point and interval estimates” (APA, 2010, p. 34).

Figure 6.4 shows for the two groups the mean differences between pre-test and post-test, for the data presented in Figure 6.3. The figure includes 95% CIs on those differences, and there are reference lines that indicate the amounts of improvement judged by the researchers to be small, medium and large, and of clinical importance. The CIs in Figure 6.4 allow us to conclude that, for the control condition, the change from pre-test to post-test is around zero, or at most small; for the treatment condition the change is of clinical importance, and likely to be large or even very large.

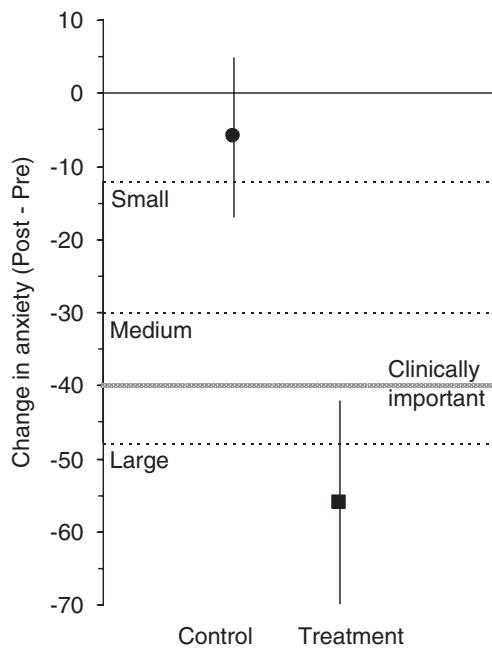


Figure 6.4 Mean Change in Anxiety Score from Pre-test to Post-test, for Treatment and Control Groups, with 95% CIs, for the Data Shown in Figure 6.3.

Dotted lines indicate reference values for changes considered small, medium and large, and the grey line the change considered clinically important.

11. Meta-analytic Thinking and Replication

Figure 6.1 shows, as we mentioned earlier, the meta-analytic combination of results from 12 studies, which might be all the available previous research. The Introduction and Discussion sections of the manuscript should both consider current research in the context of past results and likely future studies. This is meta-analytic thinking (Cumming & Finch, 2001), and it guides choice of what measures and statistics are most valuable to report, and how results are interpreted and placed in context. A forest plot (Figure 6.1) can display point and interval effect size estimates expressed in any way—as original units, or in standardized form, or as some units-free measure. For many types of research a forest plot can conveniently summarize current and past research in terms of estimation.

Replication is at the heart of science, even if in social science it has too often been neglected. All manuscript authors should consider replication, which is part of meta-analytic thinking. Even if they cannot themselves immediately run a replication, they should do all they can to assist any future replication of their work.

Replications, especially, should be judged by how well they are planned and conducted, and not by the results they obtain. Some journals are adopting the highly desirable policy of reviewing a study in advance of data collection. If the research question, and proposed design and methods, are all judged of a sufficiently high standard, the report is accepted in advance, subject only to the study being conducted as planned. Reviewers should support this enlightened policy, which should help overcome the first and third Ioannidis problems.

12. Open Materials

Providing fully detailed information about the procedure and materials used for a study is necessary for readers to understand fully what was done, and also to provide maximum assistance to any future replication efforts. The full details may need to be provided in an online supplement to a journal article, and/or in a permanent repository such as OSF. Including a protocol video, which shows how a study was actually run, can be highly valuable for anyone seeking to run a replication. In an increasing number of journals, provision of open materials, in full detail, may be acknowledged by award of the Open Materials badge.

13. Open Data

Providing open access to the full data set has many benefits: It makes meta-analysis easier, allows anyone to check for errors of analysis and interpretation, and makes it much easier for others to replicate the work. In addition, researchers can analyze the data in different ways, perhaps to address different research questions. Of course, researchers must not post sensitive or identifying information about their participants, and they need to be sure their participants have consented to anonymous data sharing, and that the intention to provide open access to data was part of the initial application for ethics approval.

Researchers might feel that, having made the enormous effort to collect their precious data, they want to be able to use it as part of future research, rather than release it immediately to other researchers. When there are good reasons, it can be acceptable to delay release of full data while the original researchers work further with it, but usually 12 months should be the maximum delay before data are made openly available. Again, the journal may offer an Open Data badge to acknowledge that the full data set is available from an open enduring repository.

Our conclusion is that it is important that authors, reviewers, and editors work together to help advance the social sciences as much as possible from the blinkered, dichotomous thinking of NHST to the richer and more informative research communication described in this chapter, and make every effort to encourage Open Science.

References

- APA. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.
- Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10, 389–396.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: Wiley.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997–1003.
- Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.
- Cumming, G. (2007). Inference by eye: Pictures of confidence intervals and thinking about levels of confidence. *Teaching Statistics*, 29, 89–93.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25, 7–29. Retrieved from tiny.cc/tnswhyhow.
- Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. *Journal of Cell Biology*, 177, 7–11.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61, 530–572.
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist*, 60, 170–180. Retrieved from tiny.cc/inferencebyeye.
- Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, 3, 299–311.
- Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York: Routledge.
- Hunt, M. (1997). *How science takes stock. The story of meta-analysis*. New York: Sage.

- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2, e124. Retrieved from www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124.
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56, 746–759.
- Kirk, R. E. (2003). The importance of effect magnitude. In S. F. Davis (Ed.), *Handbook of research methods in experimental psychology* (pp. 83–105). Malden, MA: Blackwell.
- Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366.
- Velicer, W. F., Cumming, G., Fava, J. L., Rossi, J. S., Prochaska, J. O., & Johnson, J. (2008). Theory testing using quantitative predictions of effect size. *Applied Psychology: An International Review*, 57, 589–608.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7, 632–638.
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician*, 70, 129–133.

7

Event History and Survival Analysis

Paul D. Allison

Survival analysis is a collection of statistical methods that are used to describe, explain, or predict the occurrence and timing of events. The name *survival analysis* stems from the fact that these methods were originally developed by biostatisticians to analyze the occurrence of deaths. However, these same methods are perfectly appropriate for a vast array of social phenomena including births, marriages, divorces, job terminations, promotions, arrests, migrations, and revolutions. In fact, I prefer the name *event history analysis*, widely used in the social sciences, because it more accurately captures the wide applicability of these methods. Other names include *failure time analysis*, *hazard analysis*, *transition analysis*, and *duration analysis*. In operations research, the methods are known as *reliability analysis* in reference to the reliability of machines, electronic components, and other material objects.

Although some methods of survival analysis are purely descriptive (e.g., Kaplan–Meier estimation of survival functions), most applications involve estimation of regression models, which come in a wide variety of forms. These models are typically very similar to linear or logistic regression models, except that the dependent variable is a measure of the timing or rate of event occurrence. A key feature of all methods of survival analysis is the ability to handle *right censoring*, a phenomenon that is almost always present in longitudinal event data. Right censoring occurs when some individuals do not experience any events, implying that an event time cannot be measured. Introductory treatments of survival analysis for social scientists can be found in Teachman (1983), Allison (2010, 2014), Tuma and Hannan (1984), Kiefer (1988), Blossfeld and Rohwer (2001), and Box-Steffensmeier and Jones (2004). For a biostatistical point of view, see Collett (2014), Hosmer and Lemeshow (2008), Kleinbaum and Klein (2012), or Klein and Moeschberger (2003). Specific desiderata for applied studies that use survival analysis are presented in Table 7.1 and later explained in detail.

1. Definition of the Event

The first step in any application of survival analysis is to define, operationally, the event that is to be modeled. Ideally, an event is a qualitative change that occurs at some specific, observed point in time. Classic examples include a death, a marriage, or a promotion. In such cases, where there is little ambiguity, there may be no need to explicitly define the event. Other applications may not be

Table 7.1 Desiderata for Survival Analysis.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. The event is defined in a clear and unambiguous way.	I
2. The observation period is specified with careful consideration of origin time and possible late entry.	M
3. Censoring is discussed, with indications of amount, type and reasons for censoring.	M
4. An appropriate choice is made between a discrete versus a continuous time method.	M
5. An appropriate choice is made between a parametric versus a semi-parametric method.	M
6. Choice of covariates is discussed and justified. Possible omitted covariates are considered.	M, D
7. Any time-varying covariates are appropriately defined, and a method for handling them is chosen.	M
8. If there are multiple events per individual, an appropriate method is chosen to handle the possible dependence among those events.	M
9. If there are competing risks, an appropriate method is chosen and appropriate tests are reported.	M
10. Sampling method and sample size are explained and justified.	M
11. The treatment of missing data is addressed.	M, R
12. The name and version of the software package is reported.	M, R
13. Summary statistics of measured variables are presented; information on how to gain access to the data is provided.	R
14. Graphs of the survivor function(s) are presented.	R
15. The proportional hazards (or equivalent) assumption is evaluated.	R
16. For evaluating different models, comparisons are made using statistical tests (for nested models) or information criteria (for non-nested models).	R
17. Coefficients (or hazard ratios) are reported, together with standard errors, confidence intervals and <i>p</i> -values.	R
18. Conditional survivor and/or hazard functions may be presented.	R
19. Potential methodological limitations are discussed.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

so clear cut, however. Some changes (e.g., menopause) take a while to “occur,” so it is necessary to make decisions about criteria for determining the timing of the event. It is also possible to define events with respect to quantitative variables, especially if they undergo sharp, sudden changes. For instance, a “stock market crash” could be said to occur if a particular market index falls more than 30% during a single week. Clearly, this definition involves some arbitrary choices that must be carefully considered and justified. A person could be said to “fall into poverty” if his or her income falls below some specified threshold. But this demands a rationale for choosing that threshold.

Another decision that must be made is whether to treat all events the same or to distinguish different types of events. If the event is an arrest, for example, one could either treat all arrests the same or distinguish between arrests for misdemeanors and arrests for felonies. All deaths could be treated alike, or one could distinguish between different kinds of deaths according to reported causes. Of course, such distinctions are only possible if data are available to differentiate the event types. Why do it? Usually, it is done because there are reasons to believe that predictor variables have different effects on different types of events. In such cases, the prevailing strategy is to

estimate competing risks models (see Desideratum 9). The downside of distinguishing different event types is that fewer events are available to estimate each set of parameters, which might substantially reduce statistical power.

Lastly, when individuals can have more than one event (of the same kind), one must decide whether to focus on a single (usually the first) event for each individual, or to use a method that incorporates all the repeated events. If the average number of events per individual is small, say, less than two, it is usually better to restrict attention to the first event.

2. Observation Period

Survival analysis requires that each individual be observed over some defined interval of time; if events occurred during that interval, their times are recorded. If events are not repeatable, observation is often terminated at the occurrence of an event. Decisions about starting and stopping times for the observation period should be reported and justified.

Most methods of survival analysis (e.g., Cox regression) require that the event time be measured with respect to some *origin time*. The choice of origin time is substantively important because it implies that the risk of the event varies as a function of time since that origin. In many cases, the choice of origin is obvious. If the event is a divorce, the natural origin time is the date of the marriage. In other cases, the choice is not so clear cut. If the event is a retirement, do you model age at retirement or time in the labor force?

Ideally, the origin time is the same as the time at which observation begins, and most software programs for survival analysis presume that this is the case. Frequently, however, observation does not begin until some time after the origin time. For example, although we may use date of marriage as the origin time in a study of divorce, couples may not be recruited into the study until years later. This is called *late entry* or *left truncation*. Because individuals are not at risk of an observed event until observation begins, special methods are necessary to take this into account. For more details, see Allison (2010, pp. 183–186).

3. Censoring

Censoring is endemic to survival analysis data, and any report of a survival analysis should discuss the types, causes, and treatment of censoring. By far the most common type of censoring is *right censoring*, which occurs when observation is terminated before an individual experiences an event. For example, in a study of divorce, couples that do not divorce during the observation period are right censored. All methods of survival analysis are designed to handle right censoring, and it is essential to incorporate the right censored observations in the analysis.

Standard methods for dealing with right censoring presume that such censoring is *non-informative*. Roughly speaking, that means that the fact that an individual is censored at particular point in time does not tell us anything about that individual's risk of the event. That assumption is necessarily satisfied if the censoring time (or potential censoring time) is the same for everyone in the sample. This would be the case, for example, if a sample of released prisoners were all followed for one year after their release, and the event of interest was a first arrest. Any persons not arrested during the one-year period would be censored, and all censoring would occur at one year.

However, if censoring occurs because individuals drop out of the study (at varying times), this censoring could potentially be informative. For example, in a study of mortality, people who are very sick could drop out before the study is completed. That could lead to biased estimates of parameters. Unfortunately, there is no test for the non-informative assumption and little that can be done to correct for bias due to its violation. One lesson, however, is that survival studies should be designed

and executed so as to minimize censoring due to dropouts. In any case, the proportion of censored cases due to dropouts should be reported.

A slightly less common type of censoring is *interval censoring*, which means that an individual is known to have an event between two points in time, but the exact time is unknown. For example, if a person reports being unmarried at wave 1 of a panel study but married at wave 2, then the marriage time is interval censored. If the intervals are regularly spaced, interval censoring can often be handled by discrete-time methods (see the next section). Although some survival analysis software can handle irregular patterns of interval censoring, most cannot.

The least common type of censoring is *left censoring*, which happens when an event is known to have occurred before some particular time, but the exact time is unknown. For example, in a study of first marriage, if a person is known only to have married before age 20, that person's marriage age is left censored. Note that the term left censoring is often used with a quite different meaning in the social science literature. In this alternative meaning, left censoring is said to occur when we begin observing an individual at some arbitrary point in time, but we do not know the origin time (i.e., how long it has been since the individual has been at risk of the event).

4. Discrete-Time vs. Continuous-Time Methods

If you know the exact times at which events occur, it is appropriate to use methods that treat time as continuous. If, on the other hand, you know only the month or the year of the event, you might be better off using discrete-time methods. One of the best indications of the need for discrete-time methods is the presence of large numbers of *ties*. A tie is said to occur if two individuals experience an event at the same recorded time. Occasionally, time is truly discrete in the sense that events can only occur at certain discrete points in time. For example, in most universities, faculty can only be promoted at the beginning of an academic year.

Most survival analysis software is designed for continuous-time data. If you want to go the discrete-time route, you must choose between a logit (logistic regression) model and a complementary log-log model. Logit is more appropriate for event times that are truly discrete, while complementary log-log is more appropriate for events that can happen at any time but are only observed to occur in discrete intervals. In practice, the choice is usually not very consequential.

Having chosen a discrete-time model, you must then choose an estimation method. Some Cox regression programs (e.g., SAS and Stata) have options for estimating either the logit or the complementary log-log model using partial likelihood estimation. But partial likelihood can be very computationally intensive for large samples with lots of ties. The alternative is to do maximum likelihood using conventional binary regression software. The trick is to break up each individual's event history into a set of distinct records, one for each unit of time in which the individual is observed, with a dependent variable coded 1 if an event occurred in that time unit, otherwise 0. One can then estimate the logit model using standard logistic regression software (Allison 1982, 2010). Many packages also have options for estimating the complementary log-log model.

5. Parametric vs. Semi-parametric Methods

By far, the most popular method for regression analysis of survival data is Cox regression, in which a model known as the proportional hazards model is estimated with the method of partial likelihood. Cox regression is sometimes described as *semi-parametric* because, although it is based on a parametric regression model, it does not make specific assumptions about the probability distribution of event times. By contrast, parametric regression models assume particular families of probability distributions, such as exponential, Weibull, Gompertz, lognormal, log-logistic, or gamma.

Although Cox regression is probably the better default method, there are two goals that are easily accomplished with parametric methods but difficult or impossible with Cox regression. First, parametric methods are much better at handling left censoring or interval censoring (especially if the intervals differ across individuals). Second, it is easy to generate predicted times to events with parametric methods, but awkward (and sometimes impossible) to do so with Cox regression. Sometimes people choose parametric methods because they worry that their data do not satisfy the proportional hazards assumption (see Desideratum 15). However, parametric models typically make assumptions that are even more restrictive than the proportional hazards assumption.

6. Covariates

Issues regarding covariates (also known as predictor variables, independent variables, regressors) are mostly the same in survival analysis as in linear regression and logistic regression (with the important exception described in Desideratum 7). Although it is desirable to provide a rationale for the inclusion of each covariate in the regression model, it is not essential. The consequences of including a variable that actually has no effect are minimal. The real danger, as with any regression analysis of observational data, comes from omitting variables that really have an effect on the outcome. This can lead to severe bias, especially if the omitted variable is moderately to strongly correlated with included variables. So any report of a survival regression should discuss the possibility of important variables that have not been included.

As with other kinds of regression, it is important to consider whether the covariates have nonlinear effects on the outcome and whether there are interactions among the covariates in their effects on the outcome. Strategies for testing and including such nonlinearities and interactions are basically the same as in linear regression, except that there are some special graphical diagnostics available for nonlinearities in Cox regression (Therneau & Grambsch, 2001). Multicollinearity is also a potential problem. Although survival analysis programs typically don't provide collinearity diagnostics, one can simply do a preliminary check with a linear regression program, while specifying the event time as the dependent variable. Because multicollinearity is all about correlations among the covariates, it is not necessary to evaluate it within the context of a survival analysis.

7. Time-Dependent Covariates

One major difference between survival regression and conventional linear regression is the possibility of time-dependent (time-varying) covariates. These are predictor variables whose values may change over the course of observation. For example, suppose that over a five-year period, information is recorded on any changes in marital status. Then, marital status (updated on a daily basis) could be used as a time-varying predictor of some other event, such as, an arrest.

Not all survival analysis methods and/or software can handle time-dependent covariates. For example, most programs for parametric survival models do not allow for time-dependent covariates (although that feature is available in recent releases of Stata). On the other hand, such variables are usually easy to incorporate into discrete-time methods based on logistic (or complementary log-log) regression. That is because each discrete time point is treated as a separate observation, so that any time-dependent covariates can be updated for each observation.

Cox regression is also well known for its ability to handle time-dependent covariates. However, there are two quite different approaches for implementing this capability in software packages. The “episode splitting” method requires that the data be configured so that there is a separate record for each interval of time during which all the covariates remain constant. The “programming statements” method expects one record per individual, with the time-varying covariates appearing as

separate variables for each time at which the variables are measured. The time-dependent covariates are then defined in programming statements prior to model specification. Properly implemented, these two methods will give identical results.

One potential issue with time-dependent covariates is that the frequency with which they are measured may not correspond to the precision with which event times are measured. For example, we may know the exact day on which a person died of a heart attack. Ideally, a time-dependent covariate, like smoking status, would also be measured on a daily basis. Instead, we may only have annual reports. Some form of imputation is necessary in such cases. The simplest and most common form of imputation is “last value carried forward,” although other methods should be considered.

One should also keep in mind that there may be several plausible ways of representing a time-dependent covariate. For example, smoking status could be coded as “person smoked on this day,” “number of days out of the last 30 in which the person smoked,” or “number of years of smoking prior to the current day,” and so forth. Decisions among the alternatives should be carefully considered, and may be based on empirical performance.

8. Repeated Events

If the data contain information on more than one event for each individual, special methods are needed to take advantage of this additional information and to deal with problems that may arise. If repeated events are observed for an individual, the standard strategy is to reset the clock to 0 each time an event occurs and treat the intervals between events as distinct observations. Thus, if a person is observed to have three arrests over a five-year interval, four observations would be created. The last observation would be a right-censored interval, extending from the third arrest until the end of the observation period.

Repeated events provide more statistical power, and also make it possible to control or adjust for unobservable variables that are constant over time. However, whenever there are multiple observations per individual, there is also likely to be statistical dependence among those observations. Unless some correction is made for this dependence, standard errors and p -values will be too low and confidence intervals will be too narrow. There are four widely available methods for repeated events that provide appropriate corrections for dependence. (1) Robust standard errors (also known as Huber–White or sandwich estimates) yield accurate standard errors and p -values, but leave coefficient estimates unchanged. (2) The method of generalized estimating equations (GEE) also gives corrected standard errors and p -values but, in addition, produces more statistically efficient coefficient estimates. (3) Random effects (mixed) models provide the same benefits as GEE, but also correct the coefficients for “heterogeneity shrinkage.” This is the tendency of coefficient estimates to be attenuated toward zero because of unobserved heterogeneity. (4) Fixed effects methods also correct for dependence and heterogeneity shrinkage. In addition, they actually control for all unchanging characteristics of the individual. For more details, see Allison (2010).

Keep in mind that many software packages do not provide all of these methods. Also, note that while fixed effects methods seem to offer the most advantages, they also come with important disadvantages. First, one cannot estimate the effects of variables that are constant over time, like sex or race, although such variables are implicitly controlled. Second, standard errors may be substantially larger because the estimates are based only on variation within individuals and not variation between.

9. Competing Risks

If a decision has been made to distinguish different kinds of events, an appropriate method must be chosen to handle the different event types. There are several different ways of doing this.

In the “traditional” competing risks approach, a separate model is specified for the occurrence of each type of event. These could be any of the models already discussed. If one has continuous time data, each of these models can be estimated separately using standard software for single kinds of events. The trick is that events other than the focal event type are treated as though the individual is censored at that point in time. For example, suppose you want to estimate Cox regression models for job terminations, while distinguishing between quittings and firings. You would estimate one model for quittings, treating firings as censored observations. Then you would estimate a model for firings, treating the quittings as censored observations.

Test statistics are available for testing whether coefficients for a particular variable are the same across event types (Allison, 2010). There are also statistics for testing whether *all* variables have the same coefficients across event types. These statistics can be helpful in determining whether it is really necessary to distinguish the event types. As noted earlier, one disadvantage of distinguishing event types is that the number of events may be small for each event type, leading to a loss of statistical power.

If event times are discrete, maximum likelihood estimation requires that models for competing risks be estimated simultaneously rather than separately. An attractive model that can be estimated with conventional software is the multinomial logit model, also known as the generalized logit model. Unfortunately, there is no comparable multinomial model for the complementary log-log specification.

A problem with the traditional approach to competing risks is that, like all standard survival analysis methods, it is based on the assumption that censoring is non-informative. And because competing events are treated as if the individual were censored, this implies that each type of event is non-informative for other types of events. For example, the fact that a person dies of heart disease at a given point in time should not give us any information about that person’s risk of dying of cancer (after adjusting for any covariates).

Because that assumption is implausible in many applications, some researchers have turned to an alternative approach to competing risks based on *cumulative incidence functions*. Several major statistical packages now have routines to implement this approach, both for descriptive methods similar to Kaplan–Meier (Marubini & Valsecchi, 1995) and for regression methods similar to Cox regression (Fine & Gray, 1999). Unfortunately, the cumulative incidence method is often billed as *the correct way* to handle competing risks when, in fact, the traditional approach still has a lot going for it. In my judgment, cumulative incidence methods can be very useful for purposes of description or prediction, but the regression coefficients should not be interpreted as causal parameters (Pintilie, 2006).

In some situations with multiple event types, a “conditional” approach may make more sense than either the traditional approach or the cumulative incidence methods (Allison, 2014). In this approach, the first step is to estimate a model for event timing without distinguishing the different event types. Then, restricting the sample to those individuals who experienced events, the second step is to estimate a binary or multinomial logit model predicting the type of event. This approach is attractive when the event types represent alternative means for achieving a single goal. For example, the event might be the purchase of a computer, and computers are distinguished by whether the operating system is Windows, Linux, or Macintosh.

10. Sampling Issues

There are three questions about sampling that should be addressed: What kind of sample is used? Are the analysis methods appropriate for the sampling method? Is the sample big enough? With regard to the first question, the ideal is a well-designed, well-executed probability sample. Nevertheless,

many survival analyses are carried out on a complete population (e.g., the 50 states in the US) or on a convenience sample (e.g., students who volunteered to participate in a study). Although others may disagree, I take the position that survival analysis—including the calculation of confidence intervals and hypothesis tests—is perfectly appropriate for analyzing a complete population. The statistical models that underlie such analyses are based on a hypothesis of inherent randomness in the phenomenon itself, and they do not require any randomization in the study design to justify the application of inferential techniques. The same argument could be made about convenience samples, although any conclusions might only apply to the sample at hand.

Regarding analysis, most survival analysis packages presume, by default, that the sample is a simple random sample. For many samples, however, there will be a need to adjust for clustering, stratification, and/or weighting. Although some packages are explicitly designed for survival analysis with complex samples (e.g., SUDAAN), conventional software can often do the job. Clustering can be accommodated by the methods described above for dependence with repeated events (although it might be difficult to adjust for both repeated events and cluster sampling). Stratification can usually be handled by including the stratification variables as covariates. Finally, most packages allow for differential weighting of observations. However, even if the sampling design involved disproportionate weights, it may not be necessary or desirable to incorporate those weights into the analysis (Winship & Radbill, 1994). This is most likely to be the case if the goal is to estimate an underlying causal model rather than some population regression function.

With regard to sample size, the most important thing to keep in mind is that censored observations contribute much less information than uncensored observations (events). Conventional wisdom has it that there should be at least five (some say 10) events for each parameter in the model, in order for maximum likelihood (or partial likelihood) estimates to have reasonably good properties. As for power considerations, there are numerous software packages and applets that will calculate power and sample size for a single dichotomous covariate. Væth and Skovlund (2004) showed how these programs can be easily extended to handle more complex regression problems. Some packages (e.g., Stata, SAS, PASS) have routines that will do power calculations for Cox regression analyses.

11. Missing Data

Reports of survival analysis should say something about the extent of missing data and the methods used to handle it. Of course, the default in virtually all survival packages is to do listwise deletion (complete case analysis). And if the proportion of cases lost to missing data is small (say, 10% or less), listwise deletion is probably the best choice. Other conventional methods, like (single) imputation or dummy variable adjustment, typically lead to biased parameter estimates, biased standard error estimates, or both.

For larger fractions of missing data, much better results can be obtained with multiple imputation (Allison, 2001). In this method, imputed values are random draws from the predictive distribution of the missing values given the observed values. Several data sets are created (typically five or more), each with slightly different imputed values. The analysis is performed on each data set using standard software. Then, using a few simple rules, the results are combined into a single set of parameter estimates, standard errors, and test statistics. Multiple imputation uses all the data to produce parameter estimates that are approximately unbiased and efficient. In calculating standard errors and test statistics, multiple imputation, unlike conventional imputation, also incorporates the inherent uncertainty about the values of the missing observations.

Although there are many stand-alone packages for doing multiple imputation, the process is much easier if the imputation is done within the same package used to do the analysis. Software for doing this is available for Stata, SAS, SPSS, and R.

Nearly all standard multiple imputation routines are based on the assumption that data are missing at random. This means, roughly, that the probability of missingness may depend on variables that are observed but might not depend on the values of the variables that are missing. Multiple imputation can be done under other assumptions, but the implementation is tricky and must be carefully tailored to each application.

For survival analysis, multiple imputation should only be done for missing values on the predictor variables. Cases that have missing values on the dependent variable should simply be deleted because conventional imputation software is not well suited for missing data on event timing and censoring. In setting up the imputation model, however, it is generally a good idea to include both the (logged) event time and the censoring indicator variable so that the relations between these variables and the predictors are adequately reproduced for the imputed variables.

12. Software

Nearly all the major statistical packages have programs for doing Cox regression and Kaplan–Meier estimation of survivor functions. And all can do discrete-time maximum likelihood estimation via logistic regression. Not all can estimate parametric regression models, however, and those that do may vary widely in their capabilities. For example, SAS can estimate parametric models with left and interval censoring but cannot handle time-dependent covariates. With Stata, it is just the reverse. Cox regression programs may also vary widely in their features and capabilities. SPSS, for instance, can handle time-dependent covariates, but its programming functions for defining those covariates are rather limited compared with SAS. As of this writing, I use three packages extensively: SAS, Stata, and the `survival` package for R. Although they vary to some degree in their capabilities, all three have a wide array of methods, functions, and options for survival analysis.

Some survival regression programs allow for the incorporation of unobserved heterogeneity or, equivalently, a random intercept in the model. In my judgment, this is a useful feature if individuals have repeated events (mentioned above) because it allows for dependence among the multiple observations. However, I caution against using this option in the more typical case of non-repeated events. In that situation, unobserved heterogeneity models are only weakly identified, and results may depend too critically on the particular specification.

13. Summary Statistics and Data Accessibility

As with other regression methods, it is good practice to report summary statistics for the predictor variables, usually their means and standard deviations. There is a potential complication, however, with time-dependent covariates. If you are using a method that requires multiple records per individual, like discrete-time maximum likelihood or Cox regression using the episode splitting method, you can simply calculate the means and standard deviations over the multiple records. On the other hand, if you are doing Cox regression with programming statements, the time-dependent covariates are generated during the estimation process and are not available for calculating descriptive statistics. In that case, I would simply report such statistics for the baseline measurements of the variables.

14. Survivor and Hazard Functions

Although not essential, it is commonplace and informative to present a graph of the survivor function, usually estimated via the Kaplan–Meier method. Such graphs are helpful in giving the reader a sense of the rates of event occurrence and censoring, and how those change over time. In some

fields, a cumulative failure graph is preferred over a survivor graph. The two graphs give the same information, however, because the failure probability is just 1 minus the survivor probability.

Even more informative than the survivor function is a graph of the estimated hazard function because it more directly quantifies the rate of event occurrence and how that rate changes over time. But the problem with the hazard function is that non-parametric estimates based on Kaplan-Meier require smoothing, and different smoothing algorithms can yield markedly different graphs. Therefore, if hazard graphs are to be presented, I recommend using the actuarial (life table) method. Although this requires an arbitrary choice of time intervals, results tend to be more stable than those produced by smoothing methods.

15. Proportional Hazards Assumption

Cox regression is based on the proportional hazards model. The proportional hazards assumption says, in essence, that the dependence of the hazard on time has the same basic shape for everyone, even as the magnitude of the hazard varies across individuals as a function of their predictor values. A crucial implication of this assumption is that predictor variables have the same effects at all points in time, that is, there are no interactions with time.

Although many researchers get very concerned about whether their data satisfy this assumption, I believe that those concerns are often excessive. If the assumption is violated for a particular predictor variable, it simply means that the coefficient for this variable represents a kind of “average” effect over the period of observation. For many applications, this may be sufficient. In some cases, however, the violations may be so severe that they lead to biases in the effects of other variables. In other cases, there may be direct interest in how the effect of a variable on the hazard changes over time.

A quick check of the proportional hazards assumption can be obtained by computing correlations between time (or some function of time) and “Schoenfeld residuals” which are calculated separately for each predictor. Non-zero correlations are evidence against the proportionality assumption. Several Cox regression software packages have an option to compute these statistics.

A more definitive check is to directly include interactions between predictors and time, which are specified as time-dependent covariates. Significant interactions indicate violation of the assumption. Fortunately, in this case the method of diagnosis is also the cure. By including the interactions, the Cox model is extended to allow for non-proportional hazards.

Another way to allow for non-proportional hazards is the method of stratification, which allows for different hazard functions for different categories of a categorical variable (like sex or marital status). This is a good method for controlling for a variable without imposing the proportional hazards assumption. But it does not yield any estimates of the effect of that variable, nor does it give a test of the proportional hazards assumption.

16. Model Comparisons

Researchers typically want to know how well their statistical models fit the data. Unfortunately, global or absolute measures of fit are generally not available for survival analysis models. Usually, the best we can do is to compare the relative fit of different models. If the models are nested (i.e., one model can be obtained from another by imposing restrictions on the parameters), likelihood ratio tests can be calculated by taking twice the positive difference in the log-likelihoods (or log partial likelihoods) for the two models. Such tests can tell you whether the more complicated model is significantly “better” than the simpler model. These tests are especially useful when estimating parametric models because some of the better-known parametric distributions are nested within the generalized gamma distribution.

If two models are not nested, informal comparisons can be accomplished with Akaike's information criterion (AIC) or Schwartz's Bayesian information criterion (SBC or BIC). These statistics "penalize" the log-likelihood for the number of covariates in the model, enabling one to validly compare models with different sets of covariates. Many software packages report one or both of these statistics, both for parametric models and for Cox regression models. Preference is given to models with lower values of these statistics, although no p -values can be calculated.

17. Reports of Coefficients and Associated Statistics

Results for Cox regression may be reported as either beta (β) coefficients or hazard ratios, which are just the exponentiated beta coefficients. Beta coefficients are more easily interpreted with respect to sign (positive, negative, or zero). However, their numerical magnitudes are difficult to interpret. Hazard ratios (which are always positive) may confuse some readers because a value of 1 means no effect. But the numerical magnitude has a more straightforward meaning: letting HR denote the hazard ratio, $100(\text{HR} - 1)\%$ is the percentage change in the hazard for a one-unit increase in the predictor. In this respect, they behave just like odds ratios in logistic regression. In the biomedical sciences, there is a clear preference for reporting hazard ratios, and this preference seems to be spreading to other fields as well.

If you report β coefficients, you should also report either standard errors or 95% confidence intervals. Because hazard ratios have asymmetric distributions, standard errors are not generally reported. Instead, the convention is to report 95% confidence intervals. It is optional but desirable to report p -values for testing the null hypothesis of no effect for each coefficient. Also desirable is a chi-square test for the null hypothesis that all coefficients are zero. Many authors ritually report the log-likelihood for each model, but this is usually not informative (unless it can be used to compare nested models).

18. Conditional Survivor or Hazard Functions

In Desideratum 14, I discussed the use of survivor or hazard functions as a descriptive device. After estimating a regression model, it is often desirable to illustrate its implications by displaying a model-based survivor function or hazard function. For example, if interest centers on the effect of some treatment, one could plot survivor functions for the treated vs. control groups in such a way that the plots embody any model assumptions (e.g., proportional hazards) and also control for any covariates in the model. If the variable of interest is quantitative, one can produce plots for several selected values of that variable, again while adjusting for any covariates.

19. Potential Methodological Limitations

Any application of statistical methods to real-world data is vulnerable to errors of one sort or another. Researchers need to be acutely aware of potential problems with their data and with the analytic methods they apply to those data. They also need to be upfront with their readers regarding any problems that they suspect could compromise their conclusions.

As noted in Desideratum 6, the most serious potential problem with survival analysis regression methods is the same as that for any other regression method applied to observational (non-experimental) data: the omission of variables (confounders) that affect the outcome and that are also correlated with the included variables. The omission of confounders can produce biases so severe that they lead to conclusions that are the exact opposite of the true state of affairs.

A problem peculiar to survival analysis is informative censoring (see Desideratum 3). Once the data are in hand, there is not much that can be done about this. But, if the number of randomly censored cases is substantial, research reports should discuss their potential impact. A sensitivity analysis can help to discern the potential direction of biases resulting from informative censoring.

Another potential danger comes from fitting an incorrect model. Some of the comparative statistics discussed in Desideratum 16 can be helpful in finding a good model. But it is also desirable to fit rather different models to the data and see if the results are consistent across models. For example, there is no good way to compare the fit of a Cox regression model with parametric gamma model. But it can be quite useful to fit both models to see if they lead to the same conclusions. If they do, well and good. If not, then your confidence in the results should be appropriately reduced.

References

- Allison, P. D. (1982). Discrete time methods for the analysis of event histories. In S. Leinhardt (Ed.), *Sociological methodology 1982* (pp. 61–98). San Francisco, CA: Jossey-Bass.
- Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.
- Allison, P. D. (2010). *Survival analysis using SAS: A practical guide* (2nd ed.). Cary, NC: SAS Institute.
- Allison, P. D. (2014). *Event history and survival analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Blossfeld, H.-P., & Rohwer, G. (2001) *Techniques of event history modeling: New approaches to causal analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Box-Steffensmeier, J., & Jones, B. (2004). *Event history modeling: A guide for social scientists*. Cambridge: Cambridge University Press.
- Collett, D. (2014). *Modeling survival data in medical research* (3rd ed.). New York: Chapman & Hall.
- Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94, 496–509.
- Hosmer, D. W., & Lemeshow, S. (2008). *Applied survival analysis: Regression modeling of time to event data* (2nd ed.). New York: John Wiley & Sons.
- Kiefer, N. M. (1988). Economic duration data and hazard functions. *Journal of Economic Literature*, 26, 646–679.
- Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data*. New York: Springer-Verlag.
- Kleinbaum, D. G., & Klein, M. (2012). *Survival analysis: A self-learning text* (3rd ed.). New York: Springer-Verlag.
- Marubini, E., & Valsecchi, M. G. (1995). *Analysing survival data from clinical trials and observational studies*. New York: John Wiley & Sons.
- Pintilie, M. (2006). *Competing risks: A practical perspective*. New York: John Wiley & Sons.
- Teachman, J. D. (1983). Analyzing social processes: Life tables and proportional hazards models. *Social Science Research*, 12, 263–301.
- Therneau, T. M., & Grambsch, P. M. (2001). *Modeling survival data: Extending the Cox model*. New York: Springer-Verlag.
- Tuma, N. B., & Hannan, M. T. (1984). *Social dynamics: Models and methods*. Orlando, FL: Academic Press.
- Væth, M., & Skovlund, E. (2004). A simple approach to power calculations in regression models. *Statistics in Medicine*, 23, 1781–1792.
- Winship, C., & Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods & Research*, 23, 230–257.

8

Factor Analysis

Exploratory and Confirmatory

Deborah L. Bandalos and Sara J. Finney

Factor analysis is a method of modeling the covariation among a set of observed variables as a function of one or more latent constructs. Here, we use the term *construct* to refer to an unobservable but theoretically defensible entity, such as intelligence, self-efficacy, or creativity. Such constructs are typically considered to be latent in the sense that they are not directly observable (see Bollen, 2002, for a more detailed discussion of latent constructs). The purpose of factor analysis is to assist researchers in identifying and/or understanding the nature of the latent constructs underlying the variables of interest. Technically, these descriptions exclude *component analysis*, which is a method for reducing the dimensionality of a set of observed variables through the creation of an optimum number of weighted composites. A major difference between factor and component analysis is that in the latter all of the variance is analyzed, whereas in factor analysis, only the shared (common) variance is analyzed. For this reason, factor analysis is sometimes referred to as *common factor analysis*. In many ways, however, component analysis is very similar to common factor analysis, and many of the desiderata for exploratory factor analysis presented here apply equally to component analysis. Given that the goal of component analysis is to explain as much observed variance as possible via the weighted composites and not, as in common factor analysis, to model the relations among variables as functions of underlying latent variables, those desiderata relating to the importance of theory for factor analysis do not necessarily apply to component analysis (see Widaman, 2007, for a detailed explanation of the conceptual and mathematical distinction between exploratory factor analysis and principal components analysis). This is because, although components may represent constructs, component analysis can still have utility as a data reduction method even if the components themselves are not interpreted. In such cases, the components do not provide an explanation for the variables' shared variance, but are instead used to represent that shared variance in the most parsimonious manner possible.

Two broad classes of factor analytic methods are described in this chapter: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Although these two methods both model the observed covariation among variables as a function of latent constructs, the purpose of EFA is typically to identify the latent constructs or to generate hypotheses about their possible structures, whereas the purpose of CFA is to evaluate hypothesized structures of the latent constructs and/or to develop a better understanding of such structures. CFA is a specific form of structural equation

modeling (SEM; see Chapter 33, this volume). Whereas EFA can be carried out using conventional statistics software such as SPSS and SAS, CFA requires the use of specialized software such as AMOS (Arbuckle, 2003), EQS (Bentler, 2006), LISREL (Jöreskog & Sörbom, 2006), or Mplus (Muthén & Muthén, 1998–2007).

For more in-depth treatment of exploratory factor analysis we recommend texts by Gorsuch (1983), Comrey and Lee (1992), McDonald (1985), Mulaik (2010), and Pett, Lackey, and Sullivan (2003). More in-depth treatments of CFA can be found in SEM textbooks such as those by Byrne (1998, 2001, 2006), Bollen (1989), Kline (2015), and Loehlin (2004). Brown (2015) devotes a complete text to the use of confirmatory factor analysis for applied research. Specific desiderata are provided for EFA in Table 8.1a and for CFA in Table 8.1b, and are elucidated in the subsequent sections of this chapter.

Table 8.1a Desiderata for Exploratory Factor Analysis.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
E1. Theory and/or previous research supporting the construct(s) under investigation are synthesized.	I
E2. Exploratory vs. confirmatory analysis is justified.	I, M
E3. Measured variables that operationalize the construct are presented and thoroughly justified in terms of both quantity and content. (Here, information on reliability and validity, if appropriate, is included.)	I, M
E4. Sampling method(s) and sample size(s) are discussed and justified.	M
E5. Data are screened for outliers and the method of handling missing data is discussed. The correlation matrix is presented and/or measures of amenability of data to factoring are presented and discussed. Summary statistics of measured variables are presented.	M, R
E6. Name and version of software package is reported.	M, R
E7. Method of extraction is discussed and justified.	M, R
E8. Method(s) used to determine the number of factors are discussed and decision is justified. A number of factor solutions are explored and this is clearly presented. If a model is championed over others, there is clear justification as to why this model is superior.	M, R
E9. Method of rotation is stated and justified.	M, R
E10. Justification is provided for any variables that are eliminated; model is reanalyzed after variables are eliminated.	R
E11. Parameter estimates (e.g., pattern/structure coefficients, factor correlations) are presented and discussed for final model.	R
E12. Percentages of variance accounted for (both total and by each factor) are provided and discussed.	R
E13. Appropriate interpretation of factors is provided.	R, D
E14. Factor score determinacy is evaluated	R
E15. Composite and factor score quality is evaluated using reliability and, if possible, validity evidence.	R
E16. Appropriate caveats are provided; importance of replication and limitations of the study are discussed.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

Table 8.1b Desiderata for Confirmatory Factor Analysis.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
C1. Theory and/or previous research informing the model(s) under investigation are synthesized; a set of <i>a priori</i> specified competing models, represented by path diagrams, is preferred.	I
C2. Confirmatory versus exploratory analysis is justified.	I, M
C3. Measured variables that operationalize the construct are thoroughly presented and their relations with the factor are explained. (Here, information on reliability and validity, if available, is included.)	I, M
C4. Sampling method(s) and sample size(s) are discussed and justified. The method of handling missing data is discussed.	M
C5. Data are screened for outliers and non-normality. Summary statistics of measured variables are presented.	M, R
C6. Method of estimation is discussed and justified.	M
C7. Name and version of software package is reported.	M, R
C8. Problems with model convergence, improper estimates, and/or model identification are reported and discussed.	R
C9. Recommended data-model fit indices, including standardized residuals, are presented and discussed.	R
C10. For competing models, comparisons are made using statistical tests (for nested models) or information criteria (for non-nested models).	R
C11. Post-hoc model modifications, if made, are justified on the basis of both theoretical and statistical criteria. If cross-validation is not possible, issues with post-hoc model modification are discussed.	R, D
C12. Parameter estimates (e.g., pattern/structure coefficients, factor correlations) are presented and discussed for final model	R, D
C13. Factor quality is evaluated using construct reliability estimates, variance explained estimates and, if possible, validity evidence. Composite score quality is evaluated using observed score reliability estimates (e.g., omega, alpha).	R, D
C14. Appropriate caveats are provided; importance of replication and limitations of the study, including equivalent models, are discussed.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

E1. Theory and/or Previous Research

Many of the decisions made in EFA are, at least to some degree, subjective. Although statistically and/or mathematically based guidelines exist for some decisions (e.g., determining the number of factors), many important decisions are made on the basis of congruence with theory and/or previous research. Familiarity with the theory and research findings regarding the construct to be studied is therefore essential in EFA studies. Some may argue that a reliance on prior theory is at odds with the exploratory nature of EFA. However, even though an analysis is exploratory, it is typically based on some type of theory, however rudimentary. For example, researchers may have some ideas about the number and/or the makeup of the expected factors, although these may be quite tentative. In such situations EFA can be useful because its purpose is to explore multiple possible structures that may underlie the observed covariation. EFA does not rely on strict inferential tests, which makes concerns about inflation of Type I error rates due to testing multiple hypotheses irrelevant.

However, the lack of an inferential basis means that the researcher must rely on theory and prior research to justify the decisions made in the course of the analyses.

In some cases, researchers might argue that the latent construct or scale being analyzed is new and that no underlying theory is available. Such an argument is implausible unless one is willing to assume that the variables being analyzed were selected at random. If this is not the case then some theory, however rudimentary, must have guided the selection of the variables and this theory should be explicated to the extent possible.

E2. Exploratory vs. Confirmatory

Although exploratory and confirmatory analyses are often referred to as though they represent a dichotomy, the distinction between the two is really more a matter of degree. Certainly, it is possible to use EFA in a confirmatory manner or to use CFA in an exploratory manner. However, because the CFA model is much more restrictive than that of EFA, most experts recommend that EFA be used for situations in which minimal research has been conducted regarding the structure of the construct or measure of interest. Use of CFA in such situations often results in gross misfit of the model to the data. In situations such as these, there are innumerable ways in which the model can be respecified to improve fit, and a researcher can easily be overwhelmed by the multitude of possibilities and led astray by allowing the estimation of parameters that do not lead to the generating structure. Alternative EFA models, on the other hand, typically represent a more finite set of models that differ on criteria such as the number of factors, method of extraction, and type of rotation. Thus, the choice of models to compare tends to be much clearer in EFA. In addition, exploring many different models via EFA may uncover interesting and conceptually plausible structures that may have gone unstudied if CFA were employed.

A common situation in which one must choose between EFA and CFA is the investigation of a new set of items that have been written to measure a construct hypothesized to have several dimensions. Researchers will often claim *a priori* knowledge of the underlying structure based on the fact that the items have been written to measure specific aspects of the construct. However, in our experience, items are rarely aware of the scale for which they have been written and often fail to behave as they should; so unless there is empirical evidence to support such a claim, it is probably best to begin by conducting an EFA in such situations. If a clear, interpretable structure emerges, CFA can be employed, using an *independent* sample, to further test the structure that was championed from the exploratory analyses.

As a general guideline, EFA should be used for situations in which the variables to be analyzed are either newly developed or have not previously been analyzed together, or when the theoretical basis for the factor analysis model (i.e., number of factors, level of correlation among factors) is weak. In such situations, it is not possible to specify the model *a priori* in sufficient detail to conduct a CFA. Therefore, in our view, CFA should only be used if the structure of the variables has been previously studied using EFA with an independent source of data.

E3. Measured Variables

The number and nature of the factors are dictated by the observed variables that are analyzed. When conducting an EFA, the researcher should have an in-depth understanding of the construct under study (see Desideratum E1). This understanding should, in turn, inform the researcher's justification of how the observed variables cover the breadth of the construct. The number of variables to be factored should also be informed by the complexity of the construct. If the latent construct is believed to consist of multiple dimensions, it is imperative that several observed variables that represent these dimensions are included in the factor analysis. The author should be clear that the EFA

solution is completely dependent on the variables being factored; an expected factor will not emerge if the variables do not capture the construct adequately.

The variables can take many forms: items from a scale, subscale scores, or direct measures of subject characteristics (e.g., number of occurrences of behaviors, heart rate). Although some methodologists recommend against factoring items due to their generally low reliability and lack of a continuous scale, such analyses should not be problematic if the items have at least five scale points and are reasonably intercorrelated (see Desideratum E5 for more information on this point). If analyzing items or subscales that are the composite of a set of items, the author should present the actual items, or at least example items, whenever possible (if there are no copyright or item security issues). If the specific items or subscales have been studied previously, any reliability or validity evidence should be presented. Likewise, if direct measurements are being factored, the manner in which these variables were gathered and any prior work supporting the manner in which they were gathered should be presented. Without a clear presentation of each observed variable, it is difficult for a reader to interpret the factor solution (i.e., what do the factors represent?). Therefore, the number, type, and examples of the variables to be factored must be presented.

E4. Sampling Method(s) and Sample Size(s)

Unlike the vast majority of statistical methods with which quantitative researchers are familiar, EFA is generally used in a descriptive and exploratory, rather than inferential, manner. Many statistical packages do not provide standard errors or tests of fit for EFA, although these are now available in specialized packages.¹ Although sampling theory need not be invoked to obtain factor analytic solutions, it is still the case that results can only be generalized to samples similar to that on which the analyses have been conducted, unless previous empirical evidence exists for broader generalizations. Therefore, it is important that researchers describe the makeup of their sample in sufficient detail that readers can determine the degree to which the results might generalize to populations in which they are interested.

Although many rules of thumb have been suggested for determining an adequate sample size for EFA studies, recent studies have found that the sample size necessary to obtain accurate parameter estimates depends on characteristics of the data. The primary parameter of interest in factor analytic studies is the factor loading. There are two types of loadings, known as *structure coefficients* and *pattern coefficients*. Structure coefficients represent zero-order correlations between variables and factors. Pattern coefficients represent the unique effect of a factor on a variable, with the effects of all other factors partialled out. For situations in which there is only one factor, or in which factors are uncorrelated, structure and pattern coefficients are equivalent. In other cases, however, most experts recommend taking both coefficients into account when interpreting factors (although see Mulaik, 2010 for a counterargument).

The sample size needed to obtain accurate estimates of pattern and structure coefficients depends on the level of *communality*, or amount of variance in the variables that is accounted for by the factor solution, the number of variables per factor, and the interactions of these two conditions. Specifically, samples of 100 may be sufficient to obtain accurate estimates with three factors measured by three to four variables each if communalities are at least .7, but if communalities are lower than .5, sample sizes of at least 300 would be needed. With more factors, larger samples are needed. For example, in the latter situation (communalities < .5 and three to four variables per factor), a sample size of 500 would be needed if the number of factors were increased to seven (Hogarty, Hines, Kromrey, Ferron, & Mumford, 2005; MacCallum, Widaman, Zhang, & Hong, 1999; MacCallum, Widaman, Preacher, & Hong 2001; Velicer & Fava, 1998).

E5. Data Screening

Most EFA estimation methods are not based on strict assumptions about the nature of the sample and how it was obtained. However, this does not mean that these aspects of the study can be safely ignored. Similar points can be made with regard to the nature of the variables to be analyzed and their distributions. Although use of factor analysis does not assume normality or continuousness of the variables to be analyzed, variables with non-normal distributions and/or few scale points, as well as data with outlying observations, can have substantial effects on the results of EFAs.

It is fairly well known in the factor analytic literature that variables with similar levels of skew and/or kurtosis can form artifactual factors. This occurs because variables that are similarly distributed are more highly correlated with each other than are variables with different distributions, given that all else is equal. In the educational literature, such factors are termed “difficulty factors” and occur because easy (negatively skewed) and difficult (positively skewed) achievement and aptitude items tend to form factors irrespective of item content. The presence of difficulty factors can result in solutions that are challenging to interpret at best, and misleading at worst. The severity of such problems increases with the level of non-normality. In general, if absolute skewness and kurtosis values are no greater than 2.0 (some researchers suggest a more liberal standard of 7.0 for kurtosis), little or no distortion should occur. However, it is important that researchers provide information on the distributions of the variables to be analyzed so readers are alerted to the potential for such problems. Our preference is for a table in which the mean and standard deviation of each variable, along with values of skewness and kurtosis, are reported. If the number of variables being analyzed is so large as to preclude this option, the range of values or a description of the levels of skewness and kurtosis should be provided.

On a related note, the level of measurement associated with the variables should be clearly presented to readers. Because EFA estimation is typically based on analysis of Pearson product-moment (PPM) correlation matrices, violations of the assumptions underlying PPM correlations can result in bias of EFA parameters. More specifically, continuousness of the variables and linear relations between the variables and factors are necessary to obtain accurate results. These assumptions are violated when data are dichotomous or ordinal in nature. Although there should be little bias with five or more ordered categories, for variables with fewer categories the analysis of the PPM correlation matrix can produce biased results. Researchers should note how they addressed this issue. A common solution is to use a correlation matrix that takes the categorical nature of the variables into account, such as a matrix of tetrachoric (for dichotomous scales) or polychoric (for ordinal scales) correlations (see Finney & DiStefano, 2013, for a review of this literature), coupled with a special estimator (see Desideratum C6). However, other methods (full-information; Bolt, 2005; Jöreskog & Moustaki, 2001; McLeod, Swygert, & Thissen, 2001; Moustaki, 2007; Swygert, McLeod, & Thissen, 2001) can be used as well.

Outliers can affect EFA results, just as with other analyses. For this reason, the data should be screened for multivariate outliers and the results of these analyses reported. If outliers are found, the researcher may or may not decide to delete these. Although many researchers delete outliers as a matter of course, our view is that a thoughtful analysis of the nature and possible causes of the outlying cases should precede any such decision. Such an analysis may provide information on subpopulations for which the factor model does not hold, or other potential avenues for further investigation. If the researcher ultimately decides to delete the outlying cases from the analyses, our preference is that the analyses be conducted with and without the outliers and the results of both sets of analyses reported (space permitting).

E6. Software

Commonly used software packages differ in terms of default settings, presentation of output, and even calculations (see Desideratum E12 for an example). Different versions of a given software

package may also differ in terms of these features. For this reason, researchers should provide the name and version of the package used.

E7. Method of Extraction

As noted in the introduction to this chapter, our focus is on factor analysis rather than on component analysis. However, the two types of analyses are often confused; therefore, researchers and reviewers should ensure the analysis that is most appropriate for the goals of the study has been conducted. In general, component analysis is recommended for reducing a large number of variables to a smaller, more manageable number of components, whereas factor analysis is better suited for identifying the latent constructs underlying the variable correlations. Thus, although component analysis is often used for such purposes as scale development, if the researcher's intention is to interpret the components as latent dimensions or factors then factor analysis is the more appropriate analysis.

Extraction refers to the process by which the parameters of the factor solution, which include the factor pattern coefficients and, if appropriate, structure coefficients and factor intercorrelations, are estimated. A desirable solution is one that accounts for the correlations among the variables in the sense that these correlations can be accurately reproduced as functions of the estimated parameters. However, because there is an infinite number of combinations of pattern coefficients and factor correlations that will reproduce the variable correlations equally well, there is no one method of extraction that can be considered "best" in an absolute sense. Because of this, many methods of extraction have been proposed, each with a slightly different criterion regarding what is considered "best." Within the factor analysis model, principal axis (PA) and maximum likelihood (ML) methods are commonly used. Other methods, such as generalized least squares, unweighted least squares, image analysis, and alpha factoring, are also available, although not as widely used.

Principal axis methods provide a least squares-type solution, attempting to minimize the residuals between the correlation matrix being analyzed and the matrix implied by the factor model (i.e., pattern coefficients and factor correlations). ML factor analysis explicitly takes into account the fact that a sample and not a population matrix is being analyzed, and seeks to obtain the solution that would best reproduce the population correlation values. ML factor analysis is thus an inferential method, and standard errors for model parameters as well as tests of the goodness of fit of the solution can be obtained. Although some researchers prefer ML for this reason, it should be noted that this method may not provide accurate estimates of pattern coefficients if the factors are weak and/or if the sample size is small (Briggs & MacCallum, 2003). Whichever estimation method is chosen, this choice should not be arbitrary or dictated by the defaults of the computer package used, but should be justified by the researcher.

E8. Number of Factors

A major decision in EFA studies is that of determining the number of factors to retain. Methods of determining the number of factors can be classified as statistically based, mathematically based, or more heuristic. Statistically based methods include Bartlett's test, Velicer's (1976) minimum average partial (MAP) procedure, and Horn's (1965) parallel analysis. Mathematically based methods include the eigenvalue greater than one (K1) "rule," while more heuristic methods include the scree plot and retaining the number of factors that account for a pre-specified percentage of variance. In simulation studies, the parallel analysis and MAP procedures have been found to perform best, whereas K1 consistently performs worst. However, the MAP procedure is designed to work with component, rather than factor analysis, and has not performed well with the latter analysis. Another point to keep in mind is that the simulations comparing PA, MAP, and K1 have been based

on generated data in which the factors are known to be orthogonal. With correlated factors, it is likely that methods of determining the number of factors will be less accurate. In any case, decisions regarding the number of factors to retain should never be based on one criterion alone.

Two newer methods of determining the number of factors, the Hull method (Lorenzo-Seva, 1999) and the comparison data method (Ruscio & Roche, 2012) have also shown promise. The Hull method is a heuristic technique that attempts to balance goodness-of-fit and parsimony (as operationalized by the model degrees of freedom, which are directly related to the number of factors in EFA). The comparison data method is an expansion of parallel analysis in which the generated data sets are based on the same factor structure as the data of interest, rather than being randomly generated as in PA. Both the Hull and comparison data methods have been shown to perform at least as well as PA in terms of both recovering a known number of factors and obtaining accurate estimates of loading coefficients.

Because EFA is an exploratory technique, it is expected that researchers will compare solutions with various plausible numbers of factors. Solutions with numbers of factors suggested by several of the methods mentioned previously should be examined. In making a final decision, more weight should be placed on solutions obtained from methods that have been shown empirically to perform well (such as parallel analysis, the Hull method, and the comparison data method) relative to those known to provide biased estimates of the number of factors (such as K1). In addition, solutions on which the various methods converge are typically preferred to those for which only one method provides an optimum value.

In addition to these methods, researchers should use theory and/or previous research to inform decisions regarding the number of factors. For example, if previous research or theory suggests that there should be three factors, this solution should be obtained and compared to other possible solutions, even if it is not suggested by any of the methods used in determining the number of factors. Also, factors for which only one or two variables have strong pattern/structure coefficients should be carefully scrutinized, as such factors are relatively weak and are unlikely to replicate. In situations such as this, the researcher should obtain a solution with one fewer factor to determine whether the variables can be forced onto another factor. If this does not occur, it may be the case that the variables do represent a separate factor but that there are not sufficient variables to capture it. In any case, researchers should clearly state the criteria and logic used to determine the number of factors, and justify their choice of model.

E9. Method of Rotation

Although there are many rotational methods, the primary distinction is that between those that produce orthogonal (uncorrelated) and oblique (correlated) factors. The choice between these should correspond to the researcher's theory regarding whether the dimensions of the construct being studied are correlated. In the absence of such theory, our view is that oblique rotations will generally result in more reasonable representations of the data, because the dimensions that underlie constructs in the social and behavioral sciences tend to be correlated. In addition, if an oblique rotation is applied to data in which the factors are not correlated, an orthogonal solution will result, so nothing is lost by such a specification. On the other hand, specifying an orthogonal solution when the structure is actually oblique will cause variables to load on more than one factor. This is because specification of an orthogonal solution will force any cross-factor correlations to be manifested through the factor pattern coefficients. However, keeping in mind that EFA is an exploratory procedure, it is acceptable to obtain both orthogonal and oblique rotations and compare the results. The solution that is more interpretable and theoretically justifiable can then be chosen. The important point is that the researcher must provide some justification for the rotational procedure chosen and, if choosing between different rotations, must explain the basis by which that decision was made.

Once the choice between an orthogonal and an oblique rotation has been made, the researcher has another choice among the various rotation methods in each category. In general, however, the latter choice is not as consequential as the former. Although certain rotation methods are more likely to obtain a general factor than others (e.g., quartimax), in our experience the different rotations within the orthogonal and oblique categories do not generally have a strong effect on the results. This will not always be the case, however, so researchers should provide the specific orthogonal (e.g., varimax) or oblique (e.g., oblimin, promax) rotation method used along with an explanation of why it was chosen.

E10. Variable Elimination

It is not uncommon for researchers to eliminate variables from the model on the basis of low structure, pattern, or communality values or because the variable is strongly or equally influenced by multiple factors (i.e., high multidimensionality). Although we are not against such practices in theory, we do feel these decisions are often made in a somewhat cavalier manner. The variables being analyzed were presumably chosen for some reason, and eliminating some of them changes the definition of one's construct(s) to some extent. We therefore encourage researchers to carefully consider the decision to eliminate variables from the model, with an eye toward the validity of the construct(s) being studied. Often, studies are conducted using samples that are not sufficient to support stable results. Because of this, the lack of variable saliency or problems with multidimensional variables may be the result of a lack of stability due to sampling error. In such cases, if the variables have not been analyzed together previously, our recommendation is to retain any questionable variables to determine if their transgressions are repeated upon replication of the study.

If, after careful consideration, a researcher eliminates one or more variables from the analysis, the factor model must be reanalyzed with the remaining variables. This is because the elimination of even one variable can change the factor structure. Ideally, variables should be removed one at a time and analyses rerun after each removal. It may be the case that removal of one variable eliminates problems with others. Finally, the researcher should provide a summary of the decisions made regarding variable deletions and the justifications for deletion.

E11. Parameter Values

After a factor model has been decided upon, the researcher should provide information on the model parameter estimates. There are potentially four sets of estimates authors should provide and discuss: communalities, structure coefficients, pattern coefficients, and factor correlations. As noted previously, communalities represent the proportion of a variable's total variance that is accounted for by the factor solution. Low communality estimates can be used to identify variables that are not explained well by the factor solution. If communality values are moderate to high, the structure and pattern coefficients can be interpreted to more clearly understand the relations between the factors and observed variables.

The values that represent the relations between the factors and the observed variables are often called *loadings* in EFA. Two different sets of parameter estimates are sometimes referred to as loadings: *structure coefficients* and *pattern coefficients*, making the term ambiguous. We recommend researchers use the terms structure and pattern coefficients and avoid the term loading when reporting EFA results. If an orthogonal rotation is used, or if there is only one factor, the structure and pattern coefficient estimates are equivalent and represent the simple correlation between the factor and the variable (and range between -1 and +1). Researchers should interpret these coefficients as factor-variable relations and may note that squaring these values represents the amount of variance in the variable that is explained by the factor. For oblique rotations, however, structure and

pattern coefficients are not equivalent. Thus, with correlated factors both sets of coefficients should be presented and a clear distinction between them should be made: the pattern coefficient represents the unique relations between a factor and variable, controlling for the other factors; the structure coefficient represents the simple, zero-order correlation between the factor and the variable. In other words, for obliquely rotated solutions, pattern coefficients are not correlation coefficients, but are analogous to standardized (beta) weights in multiple regression analyses (and can fall outside the range of -1 to +1). Structure coefficients are a function of the pattern coefficients and the factor correlations and represent the total effect of the factor on the variable; that is, structure coefficients represent the unique effect of a factor on a variable plus its effect via relations with other factors. A factor may have a weak unique effect on a variable (i.e., small pattern coefficient) but may have a strong total association with the variable (i.e., large structure coefficient) due to strong factor correlations. Thus, to accurately interpret the solution, the factor correlations should be reported and discussed for oblique solutions. If the factor correlations are very weak, the structure and pattern coefficients will be similar in magnitude; on the other hand, if the factor correlations are strong, the structure and pattern coefficients may be very different.

Although there are differing opinions as to which set of coefficients should be used to interpret the meaning of the factors, we agree with Gorsuch (1983, p. 207) that “The basic matrix for interpreting the factors is the factor structure. By examining which variables correlate high with the factor and which correlate low, it is possible to draw some conclusions as to the nature of the factor.” Other methodologists recommend the focus be placed on the pattern coefficients, as these values tend to indicate a clearer structure, making it easier to identify salient variables. However, one must realize that the pattern coefficients do not represent the complete relationship between the variable and the factor. Therefore, when interpreting the factor solution, we recommend attending to the structure coefficients first and then evaluating the pattern coefficients to understand the unique factor-variable relations. Both sets of coefficients, or the pattern coefficients and factor correlations (from which the structure coefficients can be computed) should be reported in a table to allow readers to best understand the factor-variable relations.

E12. Percentages of Variance

The amount of variance explained by the championed solution and by each factor should be reported. Researchers should note the variance explained by the solution and by each factor *before rotation*. Rotating the solution distributes the variance explained across the retained factors. Therefore, although the total percentage of variance explained by the solution before and after rotation remains the same, the amount of variance associated with each factor will be adjusted after rotation. These adjusted values are only calculated for orthogonal rotations and should be reported and discussed (e.g., do the retained factors explain similar amounts of variance or are certain factors associated with much more explained variance?). When using an oblique rotation the factors overlap, and the each of the correlated factors is “credited” with any shared variance in the observed variables. Because of this, the total variance explained in a variable can appear to sum to over 1.00; therefore, the percentage of variance associated with each factor is not reported for oblique solutions. Instead, one can report the sum of squared structure coefficients associated with each factor after rotation.

Researchers and reviewers should realize that the two most commonly used statistical software programs (SPSS and SAS) compute the percentage of variance explained by each factor differently for EFA (thus, the importance of noting the software used; see Desideratum E6). Both compute the percentage/proportion of variance explained when conducting an EFA, but the two packages use different denominators, resulting in values that may be very different. SPSS calculates the percentage of variance as the amount of variance explained out of the *total* variance, whereas SAS calculates

the percentage of variance explained as the amount of variance explained out of the total amount of *common* variance. Therefore, if one conducted an EFA in SPSS and SAS, the eigenvalues, communalities and parameter estimates would be the same, but the percentage of explained variance would appear larger when computed by SAS (because the denominator would be smaller). It is critical that the amount of variance explained is interpreted with this in mind. In addition, if reviewers re-analyze the data to check the results, they may obtain different values of explained variance than the author because of this software difference.

E13. Interpretation of Factors

The factor solution should be interpreted using all of the relevant information. For an oblique solution, this includes the pattern coefficients, structure coefficients, and factor correlations, whereas for an orthogonal solution pattern coefficients provide sufficient information. In addition, knowledge of the variables being factored and the theory surrounding the construct should be incorporated in the interpretation process.

Once factors are interpreted, the factor name is most often used to communicate the identity of the factor, rather than the observed variables themselves. Therefore, naming the factor is extremely important and the process of naming/interpreting the factor should be clearly communicated. Researchers should note how they used the factor-variable relations to interpret and name the factors. Interpreting the factor solution requires a determination of the value a coefficient must reach to be considered salient, or “high.” Variables with coefficients that reach this value are used to name the factor. Given the difference in the interpretation of structure and pattern coefficients, one would not expect the same value to be used for both coefficients. Common values for structure coefficients are .30 and .40. Although to some extent this choice is arbitrary, some thought should be given to choosing an appropriate value. For uncorrelated factors, squaring the structure/pattern coefficient value provides the amount of variance explained in the variable by the factor. For correlated factors, the structure coefficients are affected by the factor correlations, and thus squaring these terms yields the amount of variance in the variable that the factor can explain both uniquely and via relations with other factors. Therefore, if approximately 10% is enough shared variance to deem a variable useful for factor interpretation, then a value of .30 or .40 could be used. One practice that should be guarded against is that in which the researcher chooses the cut-off value in a self-serving manner. For instance, we have seen published articles in which a value such as .42 was chosen. Although no reason for such a choice was provided, it appeared to be motivated by the fact that use of this value would allow the researchers to ignore variables having coefficients of .41 with more than one factor.

Interpretation and naming of the factor(s) is made easier if the solution exhibits simple structure. The five criteria for simple structure as defined by Thurstone (1947) are:

1. Each variable should have at least one loading of zero.
2. Each factor should have at least f zero loadings, where f is the number of factors.
3. Every pair of factors should have several variables that load on one but not the other.
4. If there are more than four factors, every pair of factors should have several variables with zero loadings on both factors.
5. For every pair of factors, there should be only a few variables that load on both.

Although Thurstone apparently intended the first criterion to be the most important in determining simple structure, most researchers focus on the fifth; that is, on minimizing the number of

cross-loadings. Taken together, the simple structure criteria imply that each factor has strong relations with only some of the variables, which facilitates naming the factor. On the other hand, weak relations between the factor and the variable or variables that have strong relations with numerous factors make it difficult to determine what the factor represents.

It may be the case that simple structure is not achieved and this should be acknowledged and discussed. If the solution has many variables that cross-load, this may be an indication of under-factoring. On the other hand, researchers should be cautious of solutions that extract too many factors in an effort to eliminate cross-loading variables. As noted previously, factors that are represented by only one or two variables may indicate over-factoring. Such possibilities underscore the importance of examining several different solutions (see Desideratum E8). When presenting the parameter estimates (see Desideratum E11), researchers should comment on the degree to which simple structure is achieved and note any variables that contribute to the deviation from simple structure. Authors should note that variables that have strong relations with many factors are multidimensional and this finding should be discussed in the context of the theoretical conceptualization of the construct.

Finally, researchers should discuss how the factor structure (nature and number of factors) aligns with the current theoretical conceptualization of the construct. Again, it is important to note that obtained factors are completed driven by the variables that are factored (see Desideratum E3). Therefore, failure to include variables that cover the breadth of the construct will result in failure to represent the construct's "true" dimensionality. Construct coverage must be addressed by authors (i.e., this is a potential problem or can be ruled out as a problem) when discussing "unexpected" or "interesting" findings, such as obtaining a simpler factor structure than expected. Finally, authors should note that a clear understanding of the factor necessitates replication and integration of the construct into its nomological net (see Desideratum E15) because seemingly "interpretable" factors can emerge from random data.

E14. Factor Score Determinacy

In many cases, researchers are interested in computing *factor scores* to use in other analyses. For example, a researcher may want to use the factor scores as variables in an analysis such as ANOVA or regression. Factor scores are weighted sums of the standardized variables. Factor scores can be either *exact* (also called *refined*) or *approximate* (also called *coarse*). The difference is that all of the variables are used to compute exact factor scores, whereas only the variables most associated with each factor are used to compute approximate factor scores. In the discussion that follows, we refer to exact factor scores. Several types of exact factor scores can be obtained; the differences among them have to do with differences in the weights used, which in turn result in factor scores with different properties. It should be noted that different procedures for obtaining *component* scores will produce the same results. However, for common factor analysis the procedures will produce different *factor* scores.

This difference is due to the fact that exact factor scores obtained from common factor analysis are indeterminate, meaning that there is an infinite number of ways in which factor scores could be obtained that would be consistent with a given pattern or structure matrix. The reason for this indeterminacy can be seen by considering the common factor model. This model posits that the observed variables Z are functions of common factors as well as factors that are unique to each variable.

To take a simple example, it might be hypothesized that four standardized variables, z_1, z_2, z_3 , and z_4 are functions of scores on two common factors, f_1 and f_2 and four unique factors, u_1, u_2, u_3 , and u_4 :

$$\begin{aligned}z_1 &= w_{11}f_1 + w_{12}f_2 + x_1u_1 \\z_2 &= w_{21}f_1 + w_{22}f_2 + x_2u_2 \\z_3 &= w_{31}f_1 + w_{32}f_2 + x_3u_3 \\z_4 &= w_{41}f_1 + w_{42}f_2 + x_4u_4\end{aligned}$$

In these equations, w and x are weights for the common and unique factor scores in f and u . Such a system of equations is indeterminate because the number of common and unique factor scores to be estimated in f and u is greater than the number of equations, and this will be the case regardless of the numbers of variables and factors. The indeterminacy problem is not that factor scores cannot be obtained from the variable scores, it is that there are many such sets of factor scores that may be quite different from one another (see Grice, 2001, for a review of factor indeterminacy and various methods of computing factor scores). Note that a similar problem does not exist for exact scores in the component model, because there are no unique factor scores in this model.

The degree to which factor scores are indeterminate depends on the variable to factor ratio, level of communality of the variables, and the sample size, with increases in each of these leading to greater determinacy (Acito & Anderson, 1986; Gorsuch, 1983; Grice, 2001). Of these, the level of communality has been found to have the greatest effect, whereas the sample size has a relatively minor effect. We therefore recommend that measures of factor score indeterminacy be reported and discussed for situations in which factor scores are obtained from data characterized by low communalities and/or variable to factor ratios. Grice discussed several indices for evaluating the degree of indeterminacy, and provides SAS code to compute these. These indices include the multiple correlation between each factor and the variables (ρ) and the minimum possible correlation between two sets of factor scores computed in different ways ($2\rho^2 - 1$). The former index ranges from 0 to 1, with high values indicating greater degrees of determinacy. The second index ranges from -1 to +1 and represents the degree to which two sets of factor scores constructed in different ways will be correlated. Values at or below 0 for this index indicate that the two sets of scores are unrelated, or negatively correlated, thus higher values are desirable.

E15. Composite Score Quality

In this section, we discuss methods of assessing the quality of the observed composite scores (subscale scores and total scale scores) that are suggested by the results of a factor analytic study. Although factors are, theoretically, latent constructs, in practice composites of the observed variable scores are computed and used as proxies for the constructs of interest. When this is done, it is important that the quality of such composites is assessed through the appropriate reliability and validity procedures.

Although internal consistency is often of interest, the type of reliability that is most relevant for a given scale will be dictated by its purpose. In many cases, researchers compute Cronbach's coefficient alpha for the complete set of variables even though multiple subscales were suggested by the factor analysis results. For multidimensional scales, our view is that reliability estimates should be obtained for the dimensional level at which the scale will be interpreted and used. If subscales representing separate dimensions of the construct are to be used independently, reliability coefficients should be calculated for these subscales. If, however, the total scale is conceptualized and interpreted as a higher-order factor that incorporates all of the subscales, it may be useful to obtain reliability coefficients at both the total and subscale levels.

If reliability is low, this should be discussed, rather than simply reported and ignored (see Lance, Butts, & Michels, 2006, for a discussion of acceptable levels of reliability). Although it is possible to

obtain an interpretable factor structure along with low estimates of internal consistency, it is more likely that low reliability will result if the factor solution is unstable. For example, the variables being factored may have weak inter-correlations, in which case the researcher should refer readers to the variable correlations and discuss this issue. Low internal consistency can also occur when there are few variables representing a factor. In such a case, the researcher should address the coverage of the construct's breadth. Whatever the cause, it is incumbent upon researchers to provide a rationale for the credibility of their interpretation when reliability of a composite score is low.

It should also be noted that indices such as Cronbach's alpha assume the variables that make up the composite are essentially tau-equivalent. That is, variables are assumed to have equal true score variances and means that are either equal or that differ by a constant amount. When these assumptions are not met, coefficient alpha will underestimate the true level of reliability. Because the assumption of essential tau-equivalence is unlikely to be met in practice, we recommend that researchers use internal consistency indices such as coefficient omega (McDonald, 1999) that are based on less stringent assumptions.

Whenever possible, external validity evidence should be gathered to support the proposed interpretation of the factors. As noted previously, in EFA, the factors are not obtained directly, but instead are operationalized by computing either factor or composite scores. Validity evidence is therefore obtained indirectly through these scores. In what follows, we refer to the factors for simplicity, bearing in mind that it is the factor's operationalization as factor or composite scores that is actually involved in these analyses.

Because the meaningfulness of a factor is made apparent through its relations with other variables (i.e., its nomological net), this information can greatly facilitate the interpretation and naming of the factors. In fact, it is through this process that one may begin to understand what is represented by the factors (Benson, 1998). In addition, relating the factors to other variables can provide further evidence of the distinction or lack of distinction between factors (i.e., do the factors have differential predictive utility which would support their differentiation?). It is important that researchers note the validity evidence associated with the external variables; lack of such evidence severely limits the utility of these external variables in evaluating the factors under study.

E16. Caveats and Limitations

Researchers should acknowledge that the factor structure championed in the study is only one possible representation of the relations among the variables. Other models may represent the data just as well or better than the structure presented; therefore, language that suggests that this model represents "truth" or is "confirmed" should be avoided. Authors should also acknowledge the exploratory nature of the analysis. In EFA studies, it is typical to examine many solutions, and often variables are removed and data re-analyzed. This practice capitalizes on chance due to fitting the idiosyncrasies of the sample data. Therefore, authors should note that the results from the EFA represent the structure of the data for that particular sample and that there is a need for replication (cross-validation) to assess the stability of the factor structure across independent samples from the same population. Moreover, researchers should not imply that the structure championed will necessarily generalize to other populations. Further research is needed to support such generalizations.

C1. Theory and/or Previous Research

As is the case for EFA, a solid understanding of the theory underlying the latent construct being modeled is essential for CFA studies, and the points made previously with regard to EFA are equally relevant for CFA (see Desideratum E1). In fact, because CFA requires the researcher to specify the

model to be fit to the data in more detail than is the case in EFA, knowledge of theory is even more important for CFA.

Researchers should clearly specify the theoretical and/or empirical basis for the model, and, to the extent possible, should provide hypotheses about the expected sign (positive or negative) and magnitude of the coefficients to be estimated. Our preference is for researchers to present all models to be tested in the form of path diagrams (see Chapter 33, this volume, for more detail on path diagrams). Often, existing theory and research do not converge on a single plausible model. In such cases, recommended practice is to specify alternative models corresponding to different theoretical perspectives (e.g., number of factors, factor relations) *a priori*, and to evaluate these competing models against each other appropriately.

C2. Confirmatory vs. Exploratory

The distinction between exploratory and confirmatory analyses was presented previously in the context of EFA (see Desideratum E2) thus, it will not be repeated here. Instead we use this desideratum to emphasize that the ability of CFA to evaluate and compare different *a priori* specified models developed on the basis of theory is its major strength. Use of CFA should therefore be reserved for situations in which at least one theory-based model can be hypothesized. Although it is possible to use CFA in an exploratory manner, such usage often results in gross misfit of the model and it can be very difficult to uncover the structure that best represents the data through the use of CFA output. If theoretically derived *a priori* models cannot be specified, it will often be necessary to explore several different models in an attempt to determine which provides the best representation of the data; this is a task more suited to the use of EFA. As mentioned in previous desiderata, researchers should not conduct CFA to “confirm” an EFA solution using the same sample, as this practice capitalizes on chance due to fitting the idiosyncrasies of the sample data. However, EFA is often used after CFA if the *a priori* specified model(s) results in extreme misfit. Using the same sample and moving to an exploratory approach (EFA) is completely justified and recommended; if the *a priori* specified model(s) do not fit, researchers can use the data to explore the structure that does underlie the observed variables via EFA. Researchers should note that the analysis has changed from a confirmatory to an exploratory mode, clearly documenting the different models, decision processes, and appropriate caveats.

C3. Measured Variables

As noted for EFA in Desideratum E3, the observed variables under study must be clearly presented with respect to type and number, as this speaks directly to the coverage of the breadth of the construct. Information pertaining to reliability and validity should be presented if available, and, whenever possible, examples of the variables under study should be presented. An additional consideration in CFA studies is that there must be a sufficient number of variables per factor to identify the model. In general, at least three variables per factor are required, although if there are two or more correlated factors, two variables per factor can be sufficient. These guidelines pertain only to model identification. More variables are typically required to represent the scope of the constructs.

The relations between the factors and the observed variable should be clearly presented when articulating the model(s) under study (see Desideratum C1). Up until now, we have only discussed models that specify factors as the causal agents of the observed variables. In such models, the observed variables are hypothesized to be correlated with one another because they are a function of the same factor. Given this conceptualization of the factor–variable relation, the factor is deemed *latent* and the direct paths flow from the latent variable to the observed variables. It is also possible to conceptualize

a factor as being a function of the observed variables (e.g., overall stress is a function of work stress, spouse-related stress, and children-related stress). The direct paths flow from the observed variables to the *emergent* factor. A detailed description of the emergent factor model and problems surrounding its estimation can be found in Chapter 33. When discussing the measured variables, researchers should clearly present how the variables are related to the factor (we recommend a figure) and explain why the particular model chosen (latent or emergent factor) is appropriate.

C4. Sampling Method(s) and Sample Size(s)

Researchers should specify the type of sampling that was used to obtain the data. Most commonly used computer packages for structural equation modeling (SEM) analyses allow the researcher to specify sampling weights for data obtained via stratified sampling methods, and allow for nested data obtained from clustered samples. Researchers using such sampling techniques should therefore employ the proper methodology.

With regard to sample size, the guidelines presented previously in the context of EFA also apply to CFA. Because CFA is an inferential method, researchers should consider issues of power and precision in addition to accuracy of parameter estimates when deciding on the necessary sample size. Power for CFA can be computed for both individual parameter estimates and for the model as a whole (see Hancock, 2006, and Chapter 33 of this volume, for more discussion of power in SEM). Some estimation methods used with non-normally distributed and/or categorized data—such as asymptotically distribution free (ADF) and weighted least squares (WLS)—require sample sizes that are much larger than those needed for normal theory based methods.

Recent advances in missing data methodology have called into question the utility of missing data methods such as listwise and pairwise deletion. Currently, full information maximum likelihood (FIML) and multiple imputation are considered to be more acceptable methods for accommodating missing data, and most commercially available SEM software packages have incorporated at least one of these. In the FIML method, missing values are not imputed. Instead, parameter estimates are obtained from the information available from each case for the variables involved in the parameter being estimated. The cases contributing to the estimation vary across parameters because different cases may have missing values for different variables. In multiple imputation (MI), regression based methods are used to impute, or fill in the missing values using a two-step process. In the first step, missing values are imputed using stochastic regression methods. In the second step, the newly imputed data are used to create a new set of regression equations. These regression equations are used to impute a new set of data and the process continues alternating between the two steps until the desired number of imputed data sets is obtained. The analyses of interest are conducted on all the imputed data sets and the results of the analyses are combined using rules provided by Rubin (1987). Both FIML and MI assume that data are *missing at random* (MAR), meaning that the missing values for a variable do not depend on values of that variable. Researchers should identify the method used to handle missing data and address the assumptions underlying the method. In addition, the proportions of missing data across the variables in the study should be provided.

C5. Data Screening

As with EFA, the scales of the observed variables can affect results. Specifically, normal theory based methods assume that observed variables are continuous in nature and multivariately normally distributed. Variables with five or more scale points should not result in substantial bias when applying normal theory based methods, but researchers working with variables with fewer than five scale points should consider using estimation methods designed for such data (see Desideratum C6).

Unlike EFA, CFA is an inferential methodology and commonly used estimation methods in CFA, such as maximum likelihood (ML) and generalized least squares (GLS), are based on the assumption of multivariate normality of the modeled variables. Violations of this assumption can result in biased standard errors, chi-square statistic values, and approximate fit index values. Because of this, the univariate and multivariate normality of the modeled variables should be assessed and reported. Univariate skewness and kurtosis values of less than |2.0| and |7.0|, respectively, have been suggested as acceptable departures from normality. There is no generally established cutoff for multivariate kurtosis; however, values of Mardia's normalized multivariate kurtosis greater than 3.0 may produce inaccurate results when employing a normal theory estimator (Finney & DiStefano, 2013). For levels of non-normality outside these guidelines, one of the estimation methods developed for non-normally distributed data should be used (see Desideratum C6).

Screening for univariate and multivariate outliers should be conducted prior to analysis. Univariate outliers can be identified as cases with large z -scores (e.g., ± 3 standard deviations from the mean). For multivariate outliers, Mahalanobis's D or D^2 can be used. Researchers should also determine the effect of outlying cases on the parameter estimates and fit indexes. With a large sample size, these effects may be quite small. A common practice is to obtain estimates from data with and without the outliers; if results are similar, there is arguably little reason to delete outlying cases.

C6. Estimation Method

Researchers should report the type of estimation that was implemented, along with a justification for use of the chosen method. Although ML estimation is the default in virtually every SEM computer package, this method assumes data are continuous and multivariate normally distributed. Violations of either or both of these assumptions can result in biased standard errors, chi-square values, and approximate fit index values. When data depart from normality, adjustments to normal-theory standard errors and fit indices, such as the Satorra-Bentler (SB) adjustment, can be implemented. For data with fewer than five response categories, estimators such as WLS or diagonally weighted WLS should be used. WLS estimators are applied to latent correlations, which assume a normally distributed continuous variable underlies the observed categories of the variable. That is, polychoric correlations and thresholds representing the estimated cut-points between the observed categories are computed from the raw data and used in subsequent parameter estimation (Finney & DiStefano, 2013).

C7. Software

Commonly used computer packages for conducting CFA include AMOS, EQS, LISREL, Mplus, Mx, the lavaan package available in the R computing language, and SAS Proc CALIS. For basic CFA models, these packages should provide estimates that are essentially identical. However, the packages differ with regard to the estimators available, amount and type of output provided, and capabilities for advanced features such as incorporation of sampling weights, accommodation of nested data, and availability of modern missing data methodology. Researchers should therefore provide the name of the software package chosen. In addition, because software is continually upgraded, the specific version of the software package (e.g., LISREL 9.2) should always be reported.

C8. Estimation Problems

In some cases, CFA estimation can fail. Convergence failures, non-positive definite matrices, out-of-bounds parameter estimates, and relatively large standard errors signal estimation issues

(Rindskoff, 1984). Researchers should carefully examine their computer output for such problems, because problems are not necessarily flagged by software packages. It is incumbent upon the researcher to determine the cause of the problem and correct it. Failure to do so renders the parameter estimates and other statistical indexes suspect.

Estimation problems are usually the result of model misspecification, collinearity, insufficient sample size, and/or a lack of identification. In cases of model misspecification or under-identification, the model should be respecified and this change should be documented. If other steps are taken to overcome estimation problems, these should also be clearly documented. Researchers sometimes ignore or gloss over such problems, therefore reviewers and editors should be mindful that these are potentially serious issues that require explanation and remediation.

C9. Model Fit

A thorough discussion of the plethora of fit indices developed to measure the fit of CFA models to the data is beyond the scope of this chapter. Instead, we provide a general discussion of the assessment of model-data fit and focus on those indices currently recommended by methodologists. CFA models are complex, thus it is naïve to believe that model fit can be properly assessed by a single index. Most methodologists agree that fit should be assessed through the use of several different criteria. Although the chi-square test of goodness of fit has been traditionally used to assess model-data fit, many methodologists feel that it is overly stringent because (1) the null hypothesis that the model holds exactly in the population is unrealistic, and (2) the test is sensitive to sample size. Despite these shortcomings, the chi-square test should be reported, along with its degrees of freedom and p -value. Approximate fit indices should be reported as well. These indices are often categorized into three classes, which we present below along with recommended exemplars and noted “cut-off” values used to make judgments regarding the fit of the model to the data. Fit index cut-off values have been studied (e.g., Hu & Bentler, 1999) and are often reported in applied research to justify the rejection or retention of a model. These cut-off values can serve as rough guidelines, but should not be used as strict cutoffs regarding model-data fit (e.g., Marsh, Hau, & Wen, 2004).

Absolute or stand-alone indices are measures of the discrepancy between the observed sample matrix and that implied by the CFA model being tested. An example is the chi-square test of fit. Another is the standardized root mean square residual (SRMR), which is based on the average of the residuals between the observed and implied matrices. SRMR values of approximately .08 or less are considered to be indicative of good fit.

Parsimony-adjusted indices also measure the discrepancy between the observed and implied matrices, but incorporate some type of penalty for model complexity. Because data-model fit improves as parameters are added to the model, these indices evaluate the improvement in fit resulting from the additional parameters relative to the number of parameters needed to obtain this improvement. One recommended index in this class is the root mean square error of approximation (RMSEA) and its associated confidence interval. This index (or its 90% confidence limits) should be approximately .05 or below for a well-fitting model, or .08 or below for an “acceptable” model.

Incremental fit indices measure the fit of the model of interest relative to the fit of a null or baseline model. The latter model is typically one that posits no correlations among the variables. Recommended indices in this class include the comparative fit index (CFI) and the non-normed fit index (NNFI; also known as the Tucker–Lewis Index, or TLI). Values of both indices should be approximately .95 or above for a well-fitting model.

Decisions regarding data-model fit should be based on an integration of all available information. Although evaluation of global fit indices is important, the matrix of standardized covariance residuals or correlation residuals, which represents the discrepancy between corresponding elements of

the observed and model-implied matrices, should be closely inspected to identify any local areas of misfit that were masked by the global fit indices. At a minimum, the range of these residual values should be reported when discussing fit. Large residuals should be taken as indications of areas in which the model does not account for the data; they should not be ignored but should be used to diagnose and better understand model misfit. Researchers should also examine parameter estimates to determine whether the signs and magnitudes are reasonable. In addition, convergence problems or excessively large standard errors can indicate model misspecification or collinearity problems, and should be considered indications of model misfit (see Desideratum C8).

C10. Model Comparisons

CFA provides an opportunity to examine the extent to which competing models explain the inter-relationships among variables. In fact, CFA is most useful when a set of *a priori* alternative models are estimated and compared because the researcher is then able to make more informed decisions about the viability of a target model relative to competing models. That is, support for a given model should be provided by showing both acceptable fit of the model and unacceptable fit of competing models. Authors should clearly discuss the utility of testing competing models and explain the manner in which these models will be compared.

For example, items may have been constructed to scale individuals on a single construct (e.g., depression). However, in order to cover the breadth of the construct (i.e., content validity), the items span a diverse content domain. The structure of the data must be empirically assessed to evaluate if the items are indeed unidimensional, if they are multidimensional (correlated factors representing the narrow content domains), or if they are essentially unidimensional, as represented by a bifactor model. Finding that a multidimensional model fits the data well does not provide as much support for this model as finding the multidimensional model fits well in addition to its fitting better than the simpler unidimensional model and fitting just as well as the more complex bifactor model.

Competing models may be nested or non-nested, and this influences the indices used to compare models. Models are nested if the simpler model can be derived from the more complex model by fixing parameters. A chi-square difference ($\Delta\chi^2$) test can be used to compare nested models. If the $\Delta\chi^2$ is statistically significant, the model with additional parameters is inferred to better represent the data than the simpler model. A common error in CFA is to treat models with different variables or different numbers of variables as nested models. Nested models must model the same variables, as well as having nested parameters. Non-nested models can be compared, although not through the use of the $\Delta\chi^2$ test. Instead, information criteria such as the Akaike information criterion (AIC) or its rescaled versions (ECVI, CAK, CAIC), which estimate how well the model would fit in future samples (cross-validate), are often used to compare non-nested models. Models with lower values of these indices are championed over models with higher values.

Competing models should only be compared to one another if they fit well in an absolute sense, both globally and locally (see Desideratum C9). If none of the competing models are viable representations of the data, comparing them can be misleading. For example, if authors make statements such as “Model A fits better than Model B,” readers may then infer that Model A fits well in an absolute sense. Furthermore, we believe that if only one model fits the data, there is no need to compare this model with other models that do not represent the data well.

C11. Post-hoc Model Modifications

There are two forms of post-hoc model modification: the removal of nonsignificant paths from a well-fitting model (“trimming”), and the addition of paths to increase fit for a poorly fitting model.

The former involves simplifying the a priori specified model by removing paths that do not reach a particular level of statistical significance. We recommend against this practice for several reasons: (1) using the same sample to respecify and test a modified model capitalizes on sampling error and thus decreases the chance of obtaining replicable results; (2) the model no longer aligns with theory but instead is empirically based or data-driven; and (3) respecified models are often presented as though they were a priori theoretically based models, thus misleading readers as to the initial models specified and tested. If post hoc model modification is undertaken, authors should do the following at a minimum: (1) clearly present results from the a priori specified model before any paths are removed (including fit indices, parameter estimates, and standard errors); (2) present the full set of results (fit indices, all parameter estimates, and standard errors) from the modified model, making a clear statement that fit index cutoffs and *p*-values associated with the parameter estimates do not apply to modified models estimated on the same data; (3) provide a thoughtful explanation for the lack of empirical support for that path (e.g., low variability associated with the variables due to the population under study; data collection issues that impacted the variable's validity); (4) provide a clear statement regarding capitalization on chance and the possibility of lack of power (i.e., a path may not be significant because sample size was small but the same path could be found significant if a larger sample was used); and (5) call for replication given the exploratory nature of the model modification. Often researchers delete indicators that have non-significant or weak relations with their intended factors, rather than simply deleting the path from the factor to the variable. The above recommendations apply to this practice as well. However, deletion of indicators is potentially more serious because it may impact the coverage of the breadth of the construct.

Most often, model modification involves the addition of paths to the model. A priori models often do not fit the data adequately. Consequently, researchers may be tempted to add parameters to the model based on *modification indices* (MIs). MIs provide estimates of the amount by which the chi-square value would decrease if the parameter were added. A particularly egregious practice is the addition of paths between measurement error terms. These parameters are often added to models in an attempt to improve fit. However, the need for such paths indicates that the associated variables have stronger correlations than can be accounted for by the factors, which results in non-negligible correlation residuals. These residuals may be due to unmodeled method effects, similar wording effects, or other artifacts, or may indicate the need for more substantive factors. Regardless of the source, the presence of unmodeled covariation is an indication that the hypothesized structure does not hold and should be interpreted as such.

Researchers should be aware that the CFA models they specify often place very strict constraints on the parameters. Typically, CFA models represent a perfect simple structure in which variables represent one factor and have no direct relation (i.e., zero pattern coefficients) with other factors. Such models are more representative of an ideal than of reality, and it is often the case that they do not fit well. In particular, CFA models with a large number of variables often exhibit poor fit, because with more variables there are more idealized factor-variable relations of zero that may not satisfy this standard. Another way of saying this is that the model becomes more falsifiable as the number of factor-variable relations that are set to zero increases. We emphasize that we are not advocating for the relaxation of standards for fit of CFA models. Rather, we present these comments in the hope that they will motivate researchers to develop a better understanding of the potential structure of their data, thus avoiding testing one idealized model that is unlikely to represent the data and engaging in post-hoc model modification. In short, researchers are not required to specify simple structure. Instead, researchers should specify the structure that is hypothesized based on the theorized influences on the scores (e.g., item content, data collection). Models with correlated errors and complex structures (e.g., bifactor models, multitrait–multimethod models) can and should be specified a priori if that structure is thought to underlie variable relations.

Unfortunately, many researchers lose sight of the purpose of CFA, which is to allow the testing of competing *a priori* models (see Desiderata C2 and C10). If a model does not fit the data, that information, along with a diagnosis of the source of the misfit, is useful and should inform the theory under study. On the other hand, thoughtlessly modifying a model post-hoc in an attempt to make it fit the data is not the purpose of CFA and may simply lead to models that do not replicate due to fitting the idiosyncrasies of the sample data. Researchers and reviewers must keep in mind that the purpose of conducting a CFA is to gain a better understanding of the underlying structure of the variables, not to force models to fit the data. The former is a useful scientific endeavor; the latter is not.

If a model does not fit the data, we recommend that this misfit be diagnosed using standardized covariance or correlation residual values (see Desideratum C9) in conjunction with modification indices. Given the sample-specific nature of model misfit, we encourage replication studies to evaluate the stability of the misfit. If the same area of misfit is found upon replication, it should be taken seriously and possible theoretical explanations of the misfit should be presented. Given plausible and thoughtful reasons for the misfit, the model may be modified and treated as an *a priori* specified model in future studies.

Often researchers do not have multiple samples to evaluate the replicability of model misfit. In such cases, researchers may choose to add parameters suggested by modification indices using the same data on which the model was originally estimated. Although we discourage such post-hoc model modifications, if these are made, researchers should do the following at a minimum: (1) report findings (fit indices, parameters estimates, and standard errors) from the *a priori* model and the modified model; (2) provide substantively meaningful justifications for the modifications; (3) clearly note that the modifications were done post-hoc and explain the issues surrounding this practice (e.g., capitalization on chance; fit index cutoffs and *p*-values associated with parameter estimates do not apply); and (4) present the results as exploratory and call for replication of both the original model to assess the stability of model misfit and the newly proposed model. Until modified models are studied using independent samples, it is unknown if they will generalize and, in turn, if they are plausible. Thus, researchers should be very cautious when interpreting the results from modified models and avoid statements regarding the usefulness or plausibility of the model.

C12. Parameter Estimates

If a model does not fit the data adequately, the parameter estimates may be biased and should not be reported. In such cases, the focus should turn to diagnosing model misfit. When model-data fit has been deemed adequate, parameter estimates and their corresponding significance tests should be interpreted. At a minimum, the direct relations between the factors and the observed variables (pattern coefficients) should be reported in standardized form. If an observed variable serves as an indicator to only one factor, the standardized coefficient can be squared to represent variance explained in the variable by the factor. Given the variables were specifically chosen as manifestations of these factors, one would hope the variance explained (R^2) would be high (at least .50). If these values are low, implications should be discussed (see Desideratum C13).

The unstandardized coefficients, not standardized coefficients, are tested for significance. Therefore, the unstandardized coefficients should be presented with their corresponding standard errors. Reporting the unstandardized estimates and standard errors facilitates comparison of results across independent samples. There is no need to present the significance tests for parameter estimates; readers can easily compute these values if needed. Instead, a statement regarding the statistical significance of the path coefficients can be made in the text or in the table note (e.g., "All unstandardized path coefficients were significant at $p < .05$ "). However, it should be noted that

significance of the path coefficients simply means that these are significantly different from zero, not that they are, in some sense, “good” indicators of the factors. Given the variables were specifically chosen to represent the factors, statistical significance of the path coefficients would seem to be a minimum expectation. However, researchers often interpret significance as though it were evidence of the strength of the indicators.

Lastly, if one has tested a multidimensional solution with factor covariances/correlations freely estimated, these factor relations should be reported along with the corresponding significance tests. Parameter estimates can be reported in tables, on the path diagram, or a combination of the two. If several competing models fit the data well and are theoretically plausible, authors should present and interpret results from each model.

C13. Factor and Composite Score Quality

The quality of the factor and associated composite score (e.g., total score, subscale score) is assessed by the magnitude of the parameter estimates, reliability of scores, and available validity evidence. With respect to parameter estimates, given the confirmatory nature of the analyses, one expects the observed variables to relate strongly to the factor for which they serve as indicators; it is this assumption that led to the selection and use of the observed variables. Often authors conflate adequate data-model fit with strong relations between the factors and corresponding observed variables. Weak factor-variable relations can occur despite adequate model-data fit due to low observed variable correlations. If the majority of the relations between a factor and its indicators are weak ($R^2 < .5$), the author should acknowledge this and resist the temptation to label the factor as consistent with a priori expectations. Assuming the researcher posited strong relations between the variables and factors, as is typically the case, a finding of weak relations indicates that the researcher’s hypotheses about the variables are not supported. In addition, weak factor-variable relations yield low reliability estimates, which will affect external validity (relations with theoretically related constructs), making interpretation of the factor difficult.

Several indices summarizing the factor-variable relations can be examined to inform the quality of the factor (e.g., Coefficient H, percent of explained variance) and the quality of the observed composite(s) informed by the championed factor structure (e.g., omega, Cronbach’s coefficient alpha). If the intention is to model observed composite scores then the quality of these scores is of interest. Internal consistency reliability is often reported in the form of coefficient omega or alpha. Omega is similar to alpha in that it represents the proportion of variance in the observed composite score attributable to all sources of common variance. Omega differs from alpha in that omega does not assume essential tau-equivalence, thus omega can be estimated from CFA parameter estimates. OmegaH, like omega, is a model-based reliability estimate, yet unlike omega and alpha, it estimates the proportion of variance in the composite score attributed to the general factor in a bifactor CFA model (Reise, Bonifay, & Haviland, 2013). Likewise, omegaS estimates the proportion of variance in the subscale scores attributable to a specific factor in a bifactor model. OmegaH and omegaS can be extremely useful when models do not align with simple structure.

Both alpha and omega (of any form) provide estimates of reliability associated with an observed composite score. If the intention is to instead model the factor itself (i.e., structural equation model), then the quality of the factor is of interest. “Construct reliability” (Hancock & Mueller, 2001) can be conceptualized as how well the set of indicators represent the latent variable (i.e., factor). Coefficient H equals the population squared multiple correlation when regressing the construct on the indicators (i.e., proportion of construct variability explained by its indicators). Thus, unlike omega, which represents the correlation between the factor and a unit-weighted composite in order to inform the adequacy of the composite scores, Coefficient H represents the correlation between the factor and

an optimally weighted composite to inform the adequacy of the measurement model (i.e., factor) when employing these indicators. When Coefficient H is high, the latent variable is well defined by the indicators.

In addition to estimating the reliability of observed composite scores (e.g., omega, omegaH, omegaS, alpha) and factor scores (i.e., Coefficient H), researchers should assess the average amount of variance explained in the observed variables by each factor. One would hope that, on average, the factor explains at least 50% of the variance in the variables that represent the factor. The interpretation of the factor would be quite difficult if the percent of variance due to error was larger than the percent of variance attributed to the factor (Woods & Edwards, 2008). Additionally, if the variables were selected to represent a unidimensional construct yet a multidimensional solution exists in the form of a bifactor model, the percentage of the explained common variance (ECV) associated with each factor (i.e., general factor and specific factors) should be computed and interpreted. Importantly, ECV can be used to assess the strength of the general factor. The higher the ECV value for the general factor, the more unidimensional the scores (Reise, Moore, & Haviland, 2010).

If reliability is adequate, external validity evidence should be gathered. The quality of the factor is ultimately dictated by how well observed relations with other constructs align with theoretical expectations (e.g., multitrait–multimethod analysis; see Chapter 24, this volume). If authors are unable to assess these relations, the quality of the factors and what the factors represent remains in question; therefore, authors should refrain from making cavalier statements regarding the meaning of the factor and its utility until such relations have been investigated.

C14. Caveats and Limitations

Caveats and limitations for CFA are essentially the same as those for EFA (see Desideratum E16). In addition, as noted above, if post-hoc model modifications are undertaken, the authors must clearly explain the effect of this practice on the interpretation of resulting model parameters. Furthermore, researchers may be tempted to use more far-reaching language related to the plausibility of the model when employing CFA compared to EFA. Finding adequate model fit does not imply that the model represents “truth” but instead that the model is one possible representation of the structure underlying the observed variables. Moreover, researchers should focus *equally* on the adequate fit of a model and the rejection of alternative models when interpreting results. It is the combination of rejecting competing models and failing to reject a competing model that provides the most useful insight into the dimensionality of the construct under study.

Note

- 1 Although standard errors for factor analytic methods other than maximum likelihood have been derived, the computations are intensive and have not been programmed into most general-use statistical packages. However, programs such as CEFA (Browne, Cudeck, Tateneni, & Mels, 2008) and *Mplus* (Muthén & Muthén, 1998–2007) provide standard errors for coefficients.

References

- Acito, F., & Anderson, R. D. (1986). A simulation study of factor score indeterminacy. *Journal of Marketing Research*, 23, 111–118.
- Arbuckle, J. L. (2003). Amos 5 [computer software]. Chicago, IL: Smallwaters.
- Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice*, 17, 10–17, 22.
- Bentler, P. M. (2006). EQS 6.1 for Windows [computer software]. Encino, CA: Multivariate Software.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53, 605–634.

- Bolt, D. M. (2005). Limited- and full-information estimation of item response theory models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 27–71). Mahwah, NJ: Lawrence Erlbaum.
- Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research*, 38, 25–56.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York: Guilford.
- Browne, M. W., Cudeck, R., Tateneni, K., & Mels, G. (2008). CEFA: Comprehensive Exploratory Factor Analysis, version 3.00 [computer software and manual]. Retrieved from <http://faculty.psy.ohio-state.edu/browne>.
- Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS and SIMPLIS: Basic concepts, applications and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Greenwich, CT: Information Age Publishing.
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6, 430–450.
- Hancock, G. R. (2006). Power analysis in covariance structure modeling. In G. R. Hancock & R. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 69–115). Greenwich, CT: Information Age Publishing.
- Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 195–216). Lincolnwood, IL: Scientific Software International.
- Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, 65, 202–226.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185.
- Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1–55.
- Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, 36, 347–387.
- Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.80* [computer software]. Lincolnwood, IL: Scientific Software International, Inc.
- Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford.
- Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria. What did they really say? *Organizational Research Methods*, 9, 202–220.
- Loehlin, J. C. (2004). *Latent variable models* (4th ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Lorenzo-Seva, U. (1999). Promin: A method for oblique factor rotation. *Multivariate Behavioral Research*, 34, 347–365.
- MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research*, 36, 611–637.
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84–99.
- Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers to overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, 11, 320–341.
- McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.
- McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 189–216). Mahwah, NJ: Lawrence Erlbaum.
- Moustaki, I. (2007). Factor analysis and latent structure of categorical and metric data. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 293–313). Mahwah, NJ: Lawrence Erlbaum.
- Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: Taylor & Francis.
- Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.
- Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage.
- Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, 95, 129–140.
- Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, 92, 544–559.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: J. Wiley & Sons.
- Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of a known factorial structure. *Psychological Assessment*, 24, 282–292.
- Swygert, K. A., McLeod, L. D., & Thissen, D. (2001). Factor analysis for items or testlets scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 217–250). Mahwah, NJ: Lawrence Erlbaum.

- Thurstone, L. L. (1947). *Multiple factor analysis: A development and expansion of vectors of the mind*. Chicago, IL: University of Chicago Press.
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41, 321–327.
- Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, 2, 231–251.
- Widaman, K. (2007). Common factors versus components: Principals and principles, errors and misconceptions. In R. Cudeck & R. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 177–203). Mahwah, NJ: Lawrence Erlbaum.
- Woods, C. M., & Edwards, M. C. (2008). Factor analysis and related methods. In C. R. Rao, J. P. Miller, & D. C. Rao (Eds.), *Handbook of statistics, volume 27: Epidemiology and medical statistics* (pp. 367–394). Boston, MA: Elsevier North-Holland.

9

Generalizability Theory

Amy Hendrickson and Ping Yin

Generalizability theory (*G theory*) is a powerful tool in educational measurement that can help researchers and educators conceptualize better and more efficient data collection efforts based on their needs and requirements. *G theory* is a statistical theory used to assess the consistency or dependability of scores over randomly parallel replications of a measurement procedure. In order to take full advantage of the potential of generalizability theory, one must constantly ask the question of what constitutes the measurement procedure. This is such an important question in any generalizability study that almost every piece of information in the analysis will depend on it.

G theory provides a conceptual framework and a set of statistical procedures and classical test theory and analysis of variance (ANOVA) can be viewed as its parents. In *G theory*, one identifies the various sources of error in observed scores and the relations among such sources and then estimates the error variance associated with each. The analysis of any generalizability study is important because there are various choices for procedures and methods for the analysis. The strengths of *G theory* are that the relative importance of multiple sources of score variance can be estimated separately in a single analysis for a given situation as well as be used to help the decision maker to design more efficient measurement procedures for the future. Valid generalization and interpretation of *G theory* results, however, depend on a carefully conceptualized design and well-executed analysis, and the authors should document and justify each of their decisions, as outlined in this chapter.

Software packages developed specifically for generalizability analyses are GENOVA (for balanced designs), urGENOVA (for unbalanced designs primarily), and mGENOVA (for multivariate designs), all of which are available from www.education.uiowa.edu/casma/computer_programs.htm#genova. Other commercially available software packages such as SAS and SPSS may also be used for the analyses, but have limitations in their applicability.

For comprehensive descriptions of *G theory* we recommend texts by Brennan (2001), Cronbach, Gleser, Nanda, and Rajaratnam (1972), Shavelson and Webb (1991), and Traub (1994). Specific desiderata for applied studies that utilize *G theory* are presented in Table 9.1 and explained in detail subsequently.

Table 9.1 Desiderata for Generalizability Theory.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. The need for measurement of reliability over replications of a measurement procedure is made clear.	I
2. The measurement procedure is described in detail.	M
3. The objects of measurement, universe(s) of admissible observations, universe(s) of generalization, and included facets are defined to aid in the understanding of the generalizability (G) and decision (D) studies.	M
4. The design of each G and D study is described, including whether each design is univariate or multivariate and balanced or unbalanced, and including diagrams and/or tables to facilitate the understanding of each design.	M, R
5. The name and version of the utilized software package is reported.	M, R
6. Summary statistics (<i>n</i> per task, mean, SD) and variance components for all facets and their interactions are presented.	R
7. Problems with estimation are reported and discussed.	R
8. A table and discussion of G study variance components are included.	R
9. A table and discussion of D study variance components are included.	R
10. The D study error variances and generalizability and dependability coefficients are discussed and included in a table. Standard errors for variance estimates and coefficients should also be included.	R, D
11. The practical results are discussed, including (if applicable) a discussion of multivariate results and composite scores and including an evaluation of the current assessment design and of the ideal design for ensuring valid generalizations.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

1. Measurement of Reliability

Generalizability analyses are particularly appropriate when an investigator is concerned about reliability-related issues that involve generalizing over multiple tasks, raters, occasions, and so forth. Conceptually, it is useful to think of such analyses in terms of hypothetical replications of the measurement procedure. In the terminology of generalizability theory, replication is usually described in a *randomly parallel* fashion that involves different but similar instances of the measurement procedure. For example, a group of persons might be given a set of five items to complete. A randomly parallel replication of this measurement procedure would involve a different set of five items from the same item pool. The authors must discuss the need to measure variability (or consistency) over defined replications.

2. The Measurement Procedure

The authors must describe the measurement procedure, including any assessment instruments and the included tasks, as well as the administration and scoring processes. For example, for an analysis of rater and task variability over replications, the following questions should be answered: How many tasks were included? How many score points were used per task? What was the nature of the rubrics—was holistic or analytic scoring employed? How were tasks assigned to persons? How were tasks assigned to raters? How were persons assigned to raters? There should be enough detail included regarding the nature of the measurement procedure, including the assessment and administration, to support the indicated generalizability study (*G study*) design.

3. Universes, Objects of Measurement, and Facets

The *universe(s)* of admissible observations and *universe(s)* of generalization are part of the conceptual framework behind generalizability theory. The definition of these universes for a given study lends support to and aid in understanding the indicated generalizability (G) and decision (D) studies employed. As part of both types of universes, *facets*, or sets of similar conditions of measurement, must be identified and described. These facets often include raters, tasks, occasions, and so forth. The *universe(s)* of admissible observations defines the instances of the facets as well as the relations among the facets that are acceptable conditions of measurement. For example, all high school math teachers from a given state might constitute the *universe* of admissible raters, any word problem from a given pool of math items might constitute the *universe* of admissible tasks, and a pairing of any rater with any task might constitute an acceptable relationship between these facets (crossed, in this instance).

The *objects of measurement* represent the population for whom the admissible facets would be appropriate (often people or groups). Following the high school math test example above, all the students who might respond to the math items would be the objects of measurement.

While the *universe(s)* of admissible observations defines the facets and relationships as they exist in a given situation, the *universe(s)* of generalization defines the facets and relationships to which a decision-maker wants to generalize based on the results from the given situation. In this way, we can estimate the effect on score variability of modifying the numbers of and relations among the facets (e.g., raters, tasks, occasions) and thus improve future designs (for example, finding that we may use fewer raters or fewer items and still maintain acceptable levels of reliability or identifying ways to obtain more reliable results from the measurement procedure). Given a particular *universe* of generalization, a *universe score* is defined for each object of measurement in the population as their mean (or expected) score over all randomly parallel replications of the measurement procedure. Thus, the *universe score* is analogous to the true score in classical test theory. The purpose of a measurement is to accurately estimate this *universe score* based on a sample of observations.

The authors need to clearly identify and define these concepts as related to their study. Furthermore, specifics about the facets must be described, including whether they are fixed or random and crossed or nested. The identification of the facets and *universe(s)* of admissible observations and of generalization is crucial. They provide the framework for choosing G study and D study designs, and for interpreting generalizability results. Note that oftentimes the variance terms are indicated in lower case letters for G studies and in upper case letters for D study variances. Brennan (2001), in particular, encourages this practice to easily distinguish between G and D study variance components and design.

4. Study Designs

Once a population and *universe(s)* of admissible observations has been defined, the researcher collects and analyzes data to estimate the variances of the observed scores and of the facets in the *universe(s)* of admissible observations associated with the appropriate design. The design of this study, called a *generalizability study* (or *G study*), should match the facets and relationships included in the *universe(s)* of admissible observations. The variance of the facets in the *universe(s)* of admissible observations can be decomposed into several uncorrelated components based on the G study design. Each of the components is termed a *variance component*. The variance component estimates from the G study can then be used to estimate the variance components and reliability-like coefficients for one or more D studies. In any particular application of generalizability theory, there is

usually only one universe of admissible observations and one G study (i.e., a univariate design). However, there are often multiple universes of generalization and multiple decision (D) studies that are of interest to an investigator. These multiple universes and D studies might differ in terms of sample sizes for facets, which facets are fixed and which are random, and/or the structure of the D study designs.

In cases where multiple universes of admissible observations, and thus multiple universe scores for each person are of interest, a multivariate generalizability study should be employed. In a multivariate G theory design, each universe score is associated with one level of a fixed facet and for each such level there is a random effects design. These various designs are *linked* through covariance components. An example of a multivariate design is the *table of specifications* model (Brennan, 2001; Yin, 2005). In this model, a test consists of items from several content areas: a different set of items is nested within each content area, and each content area is treated as a level of a *fixed* facet. The levels are linked because the same group of persons responded to all questions for each content area.

The authors need to clearly identify and describe the design of each G and D study, including whether each design is univariate or multivariate and balanced or unbalanced. The inclusion of Venn diagrams and/or tables that outline the various variance components help to facilitate the understanding of each design. For example, see Figures 9.1 and 9.2 for Venn diagrams that depict designs in which raters are either crossed or nested within items. The distinctions between the designs are easily discernible from the diagrams.

Furthermore, the authors need to be aware of and acknowledge that the power and flexibility of generalizability theory is associated with conceptual challenges in characterizing the designs, especially those with multivariate designs. Even though there are statistical methods associated with complex generalizability designs, the identification and conception of such designs is never clear-cut.

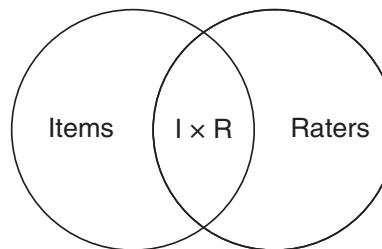


Figure 9.1 Venn Diagram Depicting Items (I) Crossed with Raters (R).

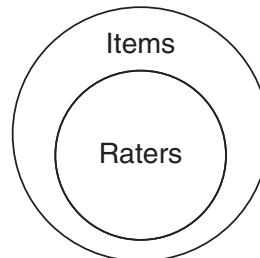


Figure 9.2 Venn Diagram Depicting Raters (R) Nested within Items (I).

5. Software

The choice of computer software for estimating G theory results depends largely on the specific method desired for estimation. There are various methods for estimating variance components in a generalizability study. For example, one type of estimation method makes normality assumptions (e.g., maximum-likelihood and restricted maximum-likelihood procedures). Another type of procedure does not require such normality assumptions. For a detailed discussion of various estimation methods and their implementation in various computer software programs, see Brennan (2001, pp. 241–247).

Relatively few software packages exist for estimating G theory results. These are the suite of GENOVA programs, SAS, SPSS, and S-Plus. These packages tend to produce similar results for most basic designs, though they have various specializations and limitations.

Considerations to keep in mind when choosing and using software packages are: the estimation method employed, memory requirements, processing time, sample size, the design implemented (e.g., unbalanced vs. balanced) and the interrelatedness of these variables. For example, maximum-likelihood (ML) and restricted maximum-likelihood (REML) procedures necessarily make heavy demands on computer resources compared to the other procedures. When using the ML or REML procedures in combination with more than 10,000 observations, memory requirements and processing time are likely to be prohibitive. Furthermore, using SPSS for G theory analyses often results in an “insufficient memory” error message even with moderately large analyses. Finally, of these programs, only urGENOVA provides G study results for unbalanced designs.

The authors should identify the software program they used, the version of the software, the estimation method employed and any problems with estimation that they encountered.

6. Summary Statistics and Variance Components

Typically, various summary statistics are presented as part of the generalizability analysis related to the particular G study design. The most important G study statistics are the estimated variance components, because they can be used to design more efficient measurement procedures and to provide information in various decision studies (e.g., obtain D study variance components and subsequent D study statistics such as error variances and generalizability coefficients). The estimated variance components for all facets presented in the G study design are typically reported as part of the summary statistics.

7. Problems with Estimation

In theory, variance components cannot be negative. In practice, however, because sampling error is likely to be present, the estimates of variance components can sometimes be negative. In particular, when the number of facets in the design is large while the sample size for a particular facet is small, a possible consequence of sampling variability related to the facet with the small sample size could result in one or more negative estimated variance components.

One obvious solution to this problem is to increase the sample size for the facet with the small sample size. However, it might not always be feasible in practice, especially when time and resources are limited. Usually, negative estimates are simply set to zero. If an investigator wants to preclude the possibility of obtaining negative estimates, Bayesian estimation procedures can be used, but they are often challenging to implement and require restrictive distributional assumptions. The negative estimate issue has theoretical implications because by definition, variance components cannot be negative. Also, because of the additive property of variance components, the negative variance

component estimates for one facet may affect the variance components for other facets and some subsequent D study statistics if using certain estimation methods such as the *expected mean square* method (see Desideratum 8). However, the negative variance component estimates seldom make much difference from a practical point of view.

The authors need to be aware that such estimation problems do occur in generalizability analyses, especially when the sample size for a facet is small. The authors should also be aware that such estimation problems do not necessarily imply inappropriate study design or estimation procedure. The best solution for the estimation problem is to increase the sample size if possible. The authors should report any estimation problems encountered in initial runs and the strategies for solving these problems, including whether any and which negative estimates were set to zero.

8. G Study Variance Components

As discussed in Desideratum 4, the main purpose of a G study is to obtain variance component estimates associated with a universe(s) of admissible observations. Using the analysis of variance method, the variance of the observed scores over the population of objects of measurement and the facets in the universe(s) of admissible observations can be decomposed into several uncorrelated components, called variance components, based on the G study design.

Several procedures can be used to obtain the estimated variance components in a G study for balanced designs, and these procedures are discussed in detail in Brennan (2001). Only a brief summary is presented in this chapter. One method is called the *expected mean square* (EMS) procedure, in which a series of EMS equations are solved by replacing parameters with estimators (Cronbach et al., 1972). The most commonly used method is called the *ANOVA method*. It can be implemented using matrix procedures or via an algorithm described in Brennan (2001). The ANOVA method has two important characteristics: (1) it does not make any normality assumptions (which are often highly suspect in generalizability theory applications), and (2) the estimated variance components are unbiased.

For unbalanced designs, two general types of methods are available to obtain estimated variance components. One type of method (i.e., maximum likelihood) assumes normality and often involves computations with large matrices, which might not be appropriate when the normality assumption is not satisfied, or when the estimation process encounters problems related to the operation with large matrices. Another type of method does not make assumptions of normality and does not typically involve operations on large matrices. This method is called *Henderson's Method 1* or the *analogous-ANOVA method* by Brennan (2001). It is difficult to mount a compelling theoretical argument for one method over another when designs are unbalanced. From a practical perspective, however, Henderson's Method 1 is relatively simple and estimated variance components can be obtained quickly.

The authors should summarize the variance component results in a table, with each G study effect and estimated variance component clearly listed. For an example of such univariate G study variance component tables, please see Brennan (2001, tables 3.3 and 3.4). Information for the estimated G study variance components is typically combined with D study results (see Desideratum 9) and presented in one table.

9. D Study Variance Components

The specification of a universe of generalization is the most important aspect of a D study. It is very common for G and D studies to have the same structure. However, designs in G and D studies do not have to be the same. For example, different D study designs can be considered based on

one G study for the purpose of designing more efficient measurement procedures and providing information for different decision studies. Therefore, D studies in generalizability analyses provide a unique and powerful tool in designing and improving measurement procedures.

The D study variance components are estimated for the universe of generalization. To obtain estimated D study variance component estimates, D study sample sizes need to be specified. Note that D study sample sizes do not need to be the same as the sample sizes for the G study. Also note that for the D study, the emphasis is on mean scores for facets considered in the design rather than individual conditions of facets.

The D study variance component for a facet can be interpreted as the variance of the distribution of mean scores for that facet, where each of these means is for the population of persons and the randomly parallel instances of the measurement procedure. Statistically, the variance of the mean score for an effect is simply the variance component for the individual effects divided by the sample size(s). Let α stand for an effect (e.g., the item effect); the D study variance component for the effect can be obtained using

$$\sigma^2(\bar{\alpha}) = \frac{\sigma^2(\alpha)}{n(\alpha)} \quad (1)$$

where $\sigma^2(\alpha)$ is the G study variance component for the effect, $\sigma^2(\bar{\alpha})$ is the D study variance component for the effect, and $n(\alpha)$ is the product of the D study sample sizes for all facets in the effect except for the object of measurement (e.g., persons). For instance, if the G study design involves persons, items, and the person \times item interaction (or residual term), the D study variance component for the person \times item interaction effect is simply the G study variance component for the person \times item interaction effect divided by the D study sample size for items.

One of the advantages of generalizability theory is the flexibility in D studies. The authors should make an effort to specify the D study sample sizes and/or D study designs with likely desired scenarios (for example, using fewer raters while still achieving the same level of reliability) so that better and more efficient studies can be designed.

10. D Study Error Variances; Generalizability and Dependability Coefficients

Two types of error variances are typically reported in a D study, *relative* and *absolute*. These error variances are very useful for making norm-referenced and criterion-referenced interpretations of scores. Specifically, the absolute error variance is the error term involved in using an individual's observed mean score as an estimate of the individual's universe score, and for random models it is obtained by summing all variance components except the universe score variance in the D study. Absolute error is often associated with domain- or criterion-referenced interpretations of scores.

The relative error variance provides information about how accurate observed deviation scores are as estimates of true deviation scores. In other words, the relative error variance summarizes how precisely examinees' observed deviations from the sample mean can predict their true deviations from the true population mean. For random models it is obtained by summing all variance components that include the universe score and at least another facet in the D study. Relative error is often associated with norm-referenced interpretations of scores.

Based on the definitions of these two error variances, which indicate that the same or more variance components are summed in the calculation of absolute error variance, absolute error variance is at least as large as relative error variance. Because the structure of the D study design often influences the magnitude of error variance, values of these absolute and relative error variances need to be considered or compared only in the context of the specific D study design.

D study coefficients are much like reliability coefficients in classical test theory, and have a range of zero to one. The two coefficients are *generalizability* and *dependability* coefficients. The generalizability coefficient is considered when making norm-referenced interpretations of scores:

$$E(\rho^2) = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)} \quad (2)$$

where $E(\rho^2)$ is used to represent a generalizability coefficient that is considered an expected squared correlation between the universe and observed scores. In equation (2), $\sigma^2(\delta)$ is the relative error variance, and $\sigma^2(p)$ is the universe score variance. Similarly, the dependability coefficient is used when criterion-referenced interpretations of scores are to be made:

$$\Phi = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\Delta)} \quad (3)$$

where Φ is the dependability coefficient, and $\sigma^2(\Delta)$ is the absolute error variance.

The generalizability coefficient appears similar to the traditional reliability coefficients defined in classical test theory (e.g., test-rest reliability, Cronbach's alpha or internal consistency reliability coefficient). However, only in generalizability theory can a researcher specifically define the universe of generalization. It is also clear that multiple estimates of generalizability coefficients can be obtained based on including different facets, D study sample sizes, and even D study designs.

It can be noted from equations (2) and (3) that the only difference between the generalizability and dependability coefficients is in the error variances. If values for the relative and absolute error variances are the same, the two coefficients are the same as well. Typically, absolute error variance is larger than relative error variance because the latter involves more variance components for the same design. Consequently, the dependability coefficient tends to be smaller than the generalizability coefficient.

When discussing different error variances and D study coefficients, the authors should be careful comparing the magnitude of the two error variances and the two D study coefficients. Depending on the G and D study designs, the two error variances and the two D study coefficients can be the same, similar, or different from each other. It is always advisable to discuss the differences instead of focusing simply on the values of these statistics.

11. Practical Results and Limitations Are Discussed

The authors should pull together the results of their D study analyses and relate them back to the particular situation they are working with and the decisions that they set out to make. They should discuss the implications of the results for their measurement procedure and provide justification for using their current procedure or rationale for changes that they expect to make to the procedure.

The authors should also have carefully interpreted the meaning of the coefficients; if the objective of the study or measurement is to rank participants, then the focus should be on the generalizability coefficient. If the observed score has utility of its own in comparison to a cut score (e.g., it is a criterion for a diagnosis, pass/fail score, or similar value), then the focus should be on the dependability coefficient. If the cut score is known, use of equation 2.54 in Brennan (2001) will almost always increase the dependability coefficient).

For all studies, but especially studies with smaller samples, authors should report standard errors for the reliability estimates and other D study results. Because Generalizability theory is a sampling theory, different samples will estimate different generalizability and dependability coefficients.

When samples are small, these can have a large range, which should be reported (Brennan, 2001, §6.3). ANOVA component estimates are known to be unstable when samples are small, and in turn can lead to incorrect conclusions. Some have called this the “Achilles’ heel” of G-theory (Brennan, 2001, pp. 210–211) and it has aroused some interest in the literature. If samples are small, authors should have to defend the stability and reliability of their estimates.

References

- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications*. Newbury Park, CA: Sage.
- Yin, P. (2005). A multivariate generalizability analysis of the Multistate Bar Examination. *Educational and Psychological Measurement*, 65, 668–686.

10

Interrater Reliability and Agreement

William T. Hoyt

When researchers make use of observer ratings (where *observer* may refer to acquaintances such as peers, family members, and teachers, or to trained observers not previously acquainted with the research participants), they provide evidence of dependability or *replicability* of ratings by reporting coefficients of reliability (for continuous scores) or agreement (for categorical ratings). Many methods have been recommended for quantifying dependability of ratings, and investigators (for whom this task is often only a peripheral issue) might not be aware of well-documented limitations of some of these approaches. Interrater reliability (for continuous rating scales) is best quantified as an *intraclass correlation coefficient* (ICC). Shrout and Fleiss (1979) provided a primer on the different types of ICCs and how to choose among them. For interrater agreement (for nominal scales) Cohen's (1960) *kappa coefficient* is recommended when there are exactly two raters, or Fleiss's (1971) extension for three or more raters. Tinsley and Weiss (1975) offered a helpful introduction to reliability and agreement, including critiques of inferior approaches to estimation. Hoyt and Melby (1999; see also Lakes & Hoyt, 2009) noted that multiple sources of error (e.g., instability of scores over time, internal inconsistency of rating scales, as well as rater variance) contribute to unreliability of ratings, and researchers may find it useful to report generalizability coefficients (Brennan, 2001; Shavelson & Webb, 1991) as a means of quantifying dependability with respect to multiple sources of error simultaneously. Schmidt and Hunter (1996; see also Schmidt, Le, & Ilies, 2003) offered a helpful discussion of the impact of measurement error on study findings, and the importance of reporting coefficients that reflect the relevant sources of error in scores. Hoyt (2000; see also Hoyt & Kerns, 1999) discussed issues for interpretation of findings in the presence of rater errors. Feldt and Brennan (1989) provided a technical treatment of the relation between reliability and generalizability coefficients.

I consider three possible contexts in which interrater reliability should be reported (Contexts A, B, C in Table 10.1). Investigators who are using established rating scales (Context A) will wish to report the dependability of scores for their own sample. Investigators who create a rating scale for a unique purpose (e.g., coding open-ended responses from participants; coding studies in a meta-analysis; Context B) will likewise wish to provide evidence that similar ratings would have been obtained using a different set of coders. Finally, investigators who are developing a new rating scale intended to be used in future substantive inquiries (Context C) should provide more detailed conceptual and psychometric information, to assist future users of this scale.

Table 10.1 Desiderata for Interrater Reliability and Agreement.

Desideratum	Context			Manuscript Section(s)*
	A	B	C	
1. Construct is clearly defined, with theory-based predictions related to establishment of reliability and validity of the measure.			●	I
2. Justification is provided for use of ratings as a source of data, with reference to the nature of the construct and past research in the area.	●	●	●	I, M
3. Procedures for generating items (or other rating instructions) are described.		●	●	M
4. Reports of interrater reliability in previous research include brief description of both targets and rater characteristics (including training procedures) in that study.	●			M
5. Rater selection and training procedures for the current study are described.	●	●	●	M
6. Procedures for computing coefficients of interrater reliability or agreement are clearly described	●	●	●	M
7. Reported interrater reliability (or generalizability) coefficients are congruent with the rating design (number of raters per target; raters crossed or not crossed with targets) used to produce the scores to be analyzed in Results section.	●	●	●	M, R
8. Dependability of ratings is appropriately considered in interpretation of findings.	●	●	●	R, D
9. Dependability of ratings based on current study leads to recommendations for appropriate use of scale in future research (e.g., rating design, rater training).			●	R, D
10. Dependability of ratings is considered in discussing study limitations and suggestions for future research. Authors acknowledge that interrater errors are one of multiple sources of error that may affect ratings data.	●	●	●	D

Note: Context A refers to investigations using established rating scales. Context B refers to “rough and ready” rating scales created uniquely for a particular study (e.g., coding systems for studies included in a meta-analysis). Context C refers to studies designed to establish reliability and validity for a new rating measure that will be used in future substantive research.

* I = Introduction, M = Methods, R = Results, and D = Discussion.

1. Definition of Construct (Context C)

It is always desirable for investigators to provide clear definitions of latent constructs for which the measured variables serve as indicators, and theoretical linkages among these constructs that lead to the research hypotheses for the study. This consideration is particularly important in instrument development studies (Context C), because the nature of the construct determines the type of validity and reliability evidence that should be sought. For example, when a construct is conceptualized as a trait, it is expected to be relatively stable over time, so that high coefficients of stability will be one characteristic of a valid measure, whereas valid measures of psychological states can be expected to have more modest stability coefficients. Thus, a careful analysis of the nature of the latent construct provides a rationale for choosing among types of evidence that bear on the validity of the scale. In the language of generalizability theory (see Chapter 9, this volume), this analysis assists the investigator in identifying the *facets of measurement* (i.e., the types of error) relevant to assessment of dependability of measurement.

Of special concern for developers of rating scales are questions about the level of inference required to judge a target’s standing on this construct, and the consistency with which this construct will be embodied in target behaviors over time and across situations. Constructs requiring low levels of inference (e.g., smiling, talking time in a conversation) can usually be reliably rated by a single rater, whereas higher level constructs (e.g., shyness, manipulativeness) will likely yield lower levels of consensus among raters. For such constructs, reliable ratings may be obtained by computing the

mean of ratings provided by multiple raters, and an important question concerns the number of raters that will be necessary to provide scores that dependably reflect the construct.

Behavioral consistency varies from construct to construct, and may also vary from person to person for a given psychological construct. For example, consensus on target extraversion tends to be substantial even after brief acquaintance with the target person, and may be adequate even among raters who view the target person in different settings. So extraversion is a medium-inference construct (i.e., there are observable behaviors that many raters agree constitute cues about a person's level of extraversion), and people may be relatively consistent about the level of extraversion they display in different situations. Many psychological constructs, however, are likely to be context-sensitive, so that observation in multiple situations would be important for validity of scores. Relatedly, behavioral indicators of some constructs (e.g., cheating) have relatively low base rates, so that valid ratings may not be obtainable without protracted periods of observation.

2. Ratings as a Source of Data (Contexts A, B, and C)

For ratings to be a source of valid information on psychological states (or traits), these states (or traits) must be reflected in observable behaviors to which the raters will have access. Investigators should provide a theory-based rationale for the types of behavioral cues to which observers have access as a basis for judgments about the construct(s) of interest. These theoretical linkages between observable behaviors and psychological characteristics can assist readers to evaluate the face validity of the rating system. For low-inference rating scales, justification of ratings as a valid data source may include theoretical explanations for typical behavioral cues (e.g., talking time in group, for ratings of extraversion) to which raters will be referred in evaluating participants' status on the rating dimension of interest. For high-inference rating scales, researchers rely on raters' global judgments of participants' status. A rationale for the use of high-inference rating scales might include theoretical considerations (e.g., an evolutionary argument that people are attuned to social cues signaling important interpersonal dimensions such as dominance and affiliation) and reference to past empirical findings that attest to the accuracy of these social perceptions.

Many types of behavioral data are subject to what Campbell, Quincy, Osserman, and Pedersen (2013) referred to as the *unitization problem*. If observed sequences of behaviors (e.g., interview data; group interactions) are to be coded in smaller chunks or *units*, investigators should explain how these units were determined (e.g., arbitrary units, such as 30-second interaction segments; or meaning units, such as speaking turns or topic switches) and the rational for this unitization scheme.

3. Item Generation (Contexts B and C)

Investigators creating a new coding system should describe how the coding instructions were created, with attention to how these operationalize the behavior-construct linkages noted in Desideratum 2. When creating a rating scale for future substantive use (Context C), investigators might want to attend in more detail to content validity of the proposed scale, perhaps including an evaluation of proposed scale items for relevance and domain coverage by experts in the area, pilot testing, or similar procedures.

4. Previously Obtained Reliability or Agreement Estimates (Context A)

Users of existing rating scales (Context A) should report evidence of interrater reliability (for continuous scales) or agreement (for nominal scales) from past studies. (Evidence of the validity of

scores on the rating scale should also be reported, as available.) Reports of dependability of ratings in past studies should be accompanied by a description of the populations under study, characteristics of the raters, and training of the raters. When multiple reliability or agreement estimates are available, it is desirable to select one from a study as similar as possible to the present rating context on these three dimensions.

One challenge that can arise regarding this desideratum is that previous users of the measure might have reported indicators of reliability or agreement that are not optimal estimators of rater consensus. For example, past users of a nominal rating scale might have quantified rater agreement as percent agreement (rather than reporting Cohen's kappa, which corrects for agreements expected due to chance). In such a case, investigators might wish to include a caveat about the limitations of available past data. To address comparability of current rater training procedures with those of the previous study, one could then report both percent agreement and the kappa coefficient for the ratings made in the present study.

5. Rater Selection and Training (Contexts A, B, and C)

By describing salient characteristics of those selected as raters, and also procedures for training those raters, investigators provide important information for future potential users of the rating scale, as well as some indication of the generalizability of these ratings. Raters in some studies are selected because of their previous acquaintance with the target person (e.g., spouse ratings, parent ratings of children), whereas in other studies raters are initially unacquainted with targets. This feature has important implications for the rating design (discussed in Desideratum 6). Unacquainted observers might be selected for their special expertise with the construct being rated (e.g., experienced clinicians versus trained undergraduate research assistants as raters of participant diagnostic category). Description of training should include approximate hours of training and brief description of training procedures (e.g., rating of standardized stimuli, criterion for determining levels of accuracy or consensus sufficient for involvement with actual study data). Ongoing monitoring and refresher training may be important to maintain rater performance in large datasets (Losada & Manolov, 2015).

6. Estimating Dependability of Ratings (Contexts A, B, and C)

Investigators should select an index of agreement (e.g., kappa, reflecting the proportion of agreement among raters corrected for chance) or reliability (e.g., the intraclass correlation coefficient, or ICC, representing the proportion of consensual variance in scores) that is appropriate to the rating scale. Coefficients of agreement are appropriate for nominal rating scales but are generally less useful for ordinal or interval scales. Coefficients of reliability are appropriate for interval or ratio scales, and are also recommended for use with ordinal scales as long as there is no reason to believe that the scale grossly violates the assumption of equal intervals between scale points. Users of interval, ratio, and most ordinal scales also have a second set of options available for quantifying dependability of ratings by reporting one or more generalizability coefficients (see Desideratum 7).

Depending on the measure being evaluated, investigators will make some decisions about how to compute a coefficient of reliability or agreement. For example, there might be a choice about the level of reliability analysis (i.e., reliability of item scores, subscale scores, or full scale scores). For some measures, scores of a single rater are expected to demonstrate adequate reliability, and only a subset of targets will be coded by a second rater for purposes of demonstrating this reliability in sample data. The research report should indicate what choices the investigator has made about these and other issues that will affect the interpretation of the reliability or agreement coefficients.

In general, it is essential to document the dependability of scores (or categorical ratings) that will actually be used in the analyses in the Results section, and computational procedures should be selected with this consideration in mind (see Desideratum 7).

When computing interrater reliability coefficients, a potentially confusing issue concerns whether the mean squares used to compute the ICC are derived from a one-way ANOVA (with targets as the lone factor) or a two-way ANOVA (including targets and raters as factors, as well as the target \times rater interaction). As Shrout and Fleiss (1979) noted, the choice of models depends on the rating design. When raters are nested within targets, the one-way ANOVA should be used. For example, if each participant is rated on trustworthiness by three acquaintances, then each target has a unique set of raters (a nested rating design), and the ICC should be computed with mean squares derived from the one-way ANOVA.

In a crossed rating design (raters crossed with targets), one set of raters evaluates all targets in the data set, and a two-way ANOVA should be used to compute the ICC. For example, if a set of 5 trained observers rates videotapes of family interactions on dimensions of cooperation and hostility, then each target is evaluated by the same set of raters (a crossed rating design), and the ICC should be computed from mean squares derived from a two-way ANOVA. When raters are crossed with targets, but the one-way ANOVA is used to derive mean squares for the ICC, reliability is generally underestimated.

Some rating studies use a “mixed” rating design. For example, the investigator might have 10 trained raters available, but will randomly assign two of these to each target. This is not a nested design because each rater judges multiple targets, but it is not crossed either because different targets are judged by different sets of raters. If Shrout and Fleiss’s (1979) formulas are used to determine interrater reliability in a mixed rating design, a one-way ANOVA should be used (similar to the nested design).

Finally, the interrater reliability coefficient estimates the proportion of observed score variance that is attributable to targets, rather than to biases of the rater(s) used in the study. It is not sufficient to report a different type of reliability coefficient (e.g., coefficient alpha, which assesses error variance attributable to differential interpretation of items) that tells nothing about rater-based errors. Further, to compute the ICC, it is necessary to have multiple targets in the reliability study (so that there will be variance in the scores attributable to the target). Although novel procedures have occasionally been recommended for computing reliability or agreement of ratings for a single target person (e.g., Cicchetti, Showalter, & Rosenheck, 1997; James, Demaree, & Wolf, 1984), these coefficients are not comparable to those that examine agreement across multiple targets (see Schmidt & Hunter, 1989 for a detailed explanation). It is the latter type of coefficient that is preferred for almost all research applications in the social and behavioral sciences.

To summarize, there are many methods to compute coefficients reflecting interrater agreement and especially interrater reliability. For readers to be able to interpret reported coefficients, it is necessary that investigators specify how the estimates were derived (full sample or subsample, rating design, model for deriving mean squares and computing the coefficient). Although such a description conveys a lot of useful information, it can usually be relatively brief. An example is provided in the next section (Desideratum 7).

7. Dependability Coefficients and Rating Design (Contexts A, B, and C)

The material in this section applies to coefficients of reliability and to coefficients of generalizability. I discuss options for computing reliability coefficients first, then briefly present an argument for considering generalizability coefficients as useful global summaries of dependability of ratings. I then present some recommendations for computing and reporting generalizability coefficients for scores involving ratings.

Intraclass correlation coefficient (ICC). Table 10.2 lists the six different ICCs discussed by Shrout and Fleiss (1979). The choice of the correct coefficient depends upon three questions. First, is the rating design nested (Case 1) or crossed (Cases 2 or 3)? Second, if the rating design is crossed, does rater variance count as error (Case 2) or not (Case 3)? Third, considering the actual scores to be analyzed (which will not necessarily be the same as those produced for the reliability study), will they be based on judgments of a single rater per target ($n'_r = 1$) or on k raters per target ($n'_r = k$)?

The first question, which concerns the distinction between nested and crossed rating designs, was discussed in Desideratum 6 above. The second question (when raters are crossed with targets) asks whether rater variance should count as error. Shrout and Fleiss (1979) related this decision to the issue of whether raters are treated as fixed (Case 3; rater variance does not count as error) or random (Case 2; rater variance counts as error).

In generalizability theory, this decision is framed in terms of the type of decision (absolute or relative) that will be based on the scores. An *absolute* decision will be based on the actual score, such as when examinees must score above a predetermined cutoff in order to pass a qualifying examination. In this case (or in any other case in which obtained ratings will be compared to ratings derived from other sets of raters or to criterion scores), rater variance counts as error, and the ICC should be computed using Shrout and Fleiss's Case 2. A *relative* decision will be based on the participant's relative standing within the sample rather than the actual score. For example, when scores derived from the ratings will be correlated with scores on another variable, only the relative standing is relevant. (The correlation coefficient would not be altered if a constant value were added to all scores in the data set.) For relative decisions, rater variance does not count as error, and Shrout and Fleiss's Case 3 should be used to compute the ICC. I return to the question of whether to count rater variance as error below, as part of the discussion of generalizability approaches.

Finally, for constructs involving little or no inference, basing each target's score on the judgment provided by a single rater might yield scores that are adequately reliable. To estimate the reliability of these scores, it will of course be necessary to have at least a subset of targets evaluated by two or more raters. Nonetheless, if the scores to be analyzed in the study are based on only a single rater, then the relevant reliability coefficient is the estimate based on $n'_r = 1$ —for example, ICC(3,1) in Table 10.2. For constructs that involve some inference on the part of raters, bias variance will be larger, and it is likely that scores based on a single rater will be relatively unreliable. To achieve adequate reliability, it will be necessary to aggregate scores across raters, using the mean or the sum of scores from multiple raters in the analysis. The reliability of this composite score will be greater than that of scores based on judgments of a single rater. Assuming that the interrater reliability analysis is conducted on data from all the raters whose data contribute to the composite scores (i.e., that $n'_r = k$), then Shrout and Fleiss's coefficient for multiple raters (e.g., ICC(3,2)) gives the correct reliability estimate.

Advantages of generalizability coefficients. The reason for reporting an interrater reliability coefficient is to inform readers (and to remind the investigator) of the proportion of replicable variance in the scores derived from the ratings, which shows the likelihood that an independent set of raters, evaluating the same target behaviors, would arrive at a similar relative ranking of the targets (ICC Cases 1 and 3) or at similar absolute scores for all targets (ICC Case 2). The complement of the reliability coefficient ($1 - \text{ICC}$) quantifies the proportion of measurement error, defined as rater error plus random error, in the scores used in the analysis.

In classical test theory, replicable variance was conceptualized as true score (valid) variance, so that the reliability coefficient estimated the proportion of score variance attributable to actual differences between targets on the construct of interest. Because at least some of the variance in observed scores is error variance, the correlation between these scores and other variables will be *attenuated* relative to (hypothetical) correlations between error-free measures of the same constructs (assuming

Table 10.2 Six Intraclass Correlation Coefficients and Comparable (One-Facet) Generalizability Coefficients.

	n^r	ICC (Shrout & Fleiss, 1979)	Generalizability Coefficient
Case 1: raters nested within targets (mean squares derived from one-way ANOVA). Use also for mixed rating design.	Single rater per target $(n'_r = 1)$	$ICC(1,1)$ $= \frac{BMS - WMS}{BMS + (k-1)WMS}$	$\hat{E}\rho^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_{r,pr,e}^2}$ $= \frac{MS_p - MS_{r,pr,e}}{MS_p + (n_r - 1)MS_{r,pr,e}}$
	Multiple raters per target $(n'_r = k)$	$ICC(1,k)$ $= \frac{BMS - WMS}{BMS}$	$\hat{E}\rho^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_{r,pr,e}^2 / n'_r}$ $= \frac{MS_p - MS_{r,pr,e}}{MS_p}$
Case 2: raters ("random") crossed with targets (mean squares derived from two-way ANOVA)	Single rater per target $(n'_r = 1)$	$ICC(2,1)$ $= \frac{BMS - EMS}{BMS + (k-1)EMS + k(EMS - EMS) / n}$	$\hat{E}\rho_{abs}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_r^2 + \hat{\sigma}_{pr,e}^2}$ $= \frac{MS_p - MS_{pr,e}}{MS_p + (n_r - 1)MS_{pr,e} + n_r(MS_r - MS_{pr,e}) / n_p}$

Multiple raters per target ($n'_r = k$)	$ICC(2,k) = \frac{BMS - EMS}{BMS + (JMS - EMS) / n}$	$\hat{E}\rho_{abs}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_r / n'_r + \hat{\sigma}_{pr,e} / n'_r}$ $= \frac{MS_p - MS_{pr,e}}{MS_p + (MS_r - MS_{pr,e}) / n_p}$
Case 3: raters (“fixed”) crossed with targets (mean squares derived from two-way ANOVA)	Single rater per target ($n'_r = 1$)	$ICC(3,1) = \frac{BMS - EMS}{BMS + (k-1)EMS}$ $= \frac{MS_p - MS_{pr,e}}{MS_p + (n_r - 1)MS_{pr,e}}$
	Multiple raters per target ($n'_r = k$)	$ICC(3,k) = \frac{BMS - EMS}{BMS}$ $= \frac{\hat{E}\rho_{abs}^2}{\hat{\sigma}_p^2 + \hat{\sigma}_{pr,e} / n'_r}$ $= \frac{MS_p - MS_{pr,e}}{MS_p}$

Note: For one-way ANOVA (Case 1), $BMS = MS_p$ = mean square between persons (targets); $WMS = MS_{rpr,e}$ = mean square error (confounds variance attributable to raters, person x rater interaction, and error). For two-way ANOVA, $BMS = MS_p$ = mean square between persons (targets); $JMS = MS_{pr,e}$ = mean square error (confounds variance attributable to person x rater interaction and error). For the ICC, k represents the number of raters per target used to estimate the ICC, whereas in G studies, this quantity is noted n_r (so $n_r = k$ for all formulas); n represents the number of targets used to estimate the ICC, whereas in G studies this quantity is noted n_p (so $n = n_p$ for all formulas); n' represents the number of raters whose scores are actually aggregated to determine participant scores, which may be different from k . Shrout & Fleiss (1979) provided formulas for computing the ICC for two possible values of n'_r : 1 and k . $\hat{E}\rho^2$ is a common notation for the generalizability coefficient, which represents the estimated expected value of the squared correlation between observed scores and universe (a.k.a. “true”) scores. When raters are crossed with persons (targets), investigators need to decide whether rater variance ($\hat{\sigma}_r^2$) counts as error (Case 2) or not (Case 3). Cronbach et al. (1972) suggested that Case 2 coefficients ($\hat{E}\rho_{abs}^2$) apply when decisions will be based on absolute scores — e.g., if ratings determine scores on a qualifying examination on which examinees must exceed a predetermined cutoff score to pass. Case 3 coefficients ($E\rho_{rel}^2$) apply when decisions will be based on relative scores (i.e., when the relative standing is important, but adding or subtracting a constant from all scores would not change the decision) — e.g., when scores will be correlated with scores on another variable to determine the degree of association between constructs.

independence of error terms for the measures of these constructs). Thus, the coefficient of reliability has important implications for interpretation of findings, in that it provides a means to estimate the magnitude of attenuation in effect size estimates (e.g., correlation coefficients, regression coefficients) derived from scores based on ratings. In many research contexts, it is useful to report effect sizes corrected for attenuation, to enhance comparability of these findings with effect sizes derived from other studies using different measurement methods.

By the 1940s, psychometricians were in widespread agreement that the idealized interpretation of reliability coefficients as estimates of the proportion of true score variance in scores represented a simplification, because of the difficulty of finding a replicated measure that is truly *parallel* (meaning that all covariance between replicated scores is attributable to true scores) in the classical sense. In practice, replicable variance in a reliability analysis usually includes variance attributable to one or more sources of error that were not varied in the reliability study.

To illustrate this principle for ratings measures, consider the case of a team of observers who rate videotaped family interactions, where the ratings will be used to derive a score on hostility for each family. The ICC for these hostility scores quantifies the extent to which raters who observe the same set of family behaviors agree in their judgments of hostility (i.e., the generalizability over raters of scores based on the same videotape), which provides an estimate of the effects of rater error on hostility scores. However, there are other sources of error that might contribute to the generalizability of ratings, depending on the use that will be made of these scores. For example, we could ask how scores derived from these videotaped interactions would compare with scores from other videotapes at a different time, or in a different setting (i.e., generalizability over occasions or settings). We might also be curious about errors attributable to the wording of items to which the raters responded (i.e., generalizability over items on the rating measure).

Because these other sources of error (occasions, settings, items) were held constant in our study of interrater reliability, any variance attributable to these sources is *replicable* variance in these studies. The ICC is an estimate of the proportion of replicable variance in ratings, but this replicable variance is not all *true score* variance (i.e., valid variance in family hostility levels). If the goal is to obtain a dependability coefficient that estimates the proportion of valid variance in ratings (and therefore can be used to estimate the degree of attenuation in effect sizes derived from these scores), then the investigator needs to think carefully about which sources of measurement error are important in this rating design. If the research hypothesis concerns the general level of hostility in family interactions, then it is important to know whether scores obtained from a given set of raters, using the specific items on the rating scale as applied to a videotaped family interaction on a single occasion, generalize to scores obtained from different raters, using different but still relevant items, applied to family interactions on different (but presumably temporally proximate) occasions.

Generalizability theory (GT; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) is a set of techniques for simultaneously evaluating generalizability (dependability) of measurement across multiple sources of error, and yields generalizability (*G*) coefficients that are interpreted as ICCs representing the proportion of valid (i.e., generalizable) variance in ratings that would be expected in a number of possible future rating designs. Because multiple sources of error are likely to be a concern for users of observer ratings, GT may be preferable to the standard approaches for determining interrater reliability. *G* coefficients indicate the proportion of replicable variance when multiple sources of error are taken into account, and can provide important guidance for improving dependability of measurement in future research applications, as well as understanding the likely impact of measurement error in attenuating study findings.

Computing and reporting G coefficients. Hendrickson and Yin (Chapter 9, this volume) describe the process of conducting *G* studies: choosing important *facets* (sources) that contribute to variability in observed scores; collecting data from multiple raters (and also multiple levels of other facets); partitioning variance in ratings into variance attributable to target, facet main effects, and interaction of

facets with target and with one another. G coefficients can then be computed for the scores that will be analyzed in the present study. Because many different G coefficients can be estimated from the same set of variance components (depending on how scores are computed), it is generally helpful to provide the formula used in computing this coefficient. In general, G coefficients are computed as a ratio of target variance (usually notated $\hat{\sigma}_p^2$ or $\text{var}(p)$) to the appropriate estimate of total variance.

Hendrickson and Yin (Chapter 9, this volume) also describe the difference between the *G study* used to provide initial psychometric data on a new measurement procedure and the *D study* (aka *decision study*) in which scores on the measure will be analyzed to make decisions about attributes of the target persons or about theoretical associations among the measured construct and other variables. For researchers who are creating rating measures to be used in future D studies (Context C in Table 10.1), GT is a particularly attractive approach, as the variance components from the G study can be used to forecast G coefficients for a variety of possible rating designs that might be used in future D studies. It can be helpful to tabulate the predicted coefficients for different designs (e.g., varying the number of raters and occasions) to illustrate how these choices on the part of future users are likely to impact dependability of scores.

Notational issues: ICCs and G coefficients. One challenge for researchers interested in using GT may be unfamiliarity with the standard notation developed by Cronbach et al. (1972) to elucidate this framework. A helpful reference point for gaining comfort with GT notation is provided in Table 10.2, where formulas derived from basic (one-facet) G study designs are compared to the equivalent formulas in the more familiar ANOVA notation employed by Shrout and Fleiss (1979). All of these coefficients are derived from one-facet G studies, because raters (r) are the only source of error that is varied in the G study. For example, Shrout and Fleiss's Case 1 corresponds to a G study where raters are nested within targets. In this design, $\text{var}(r)$, $\text{var}(pr)$, and $\text{var}(e)$ are confounded and cannot be decomposed into separate variance estimates. This is indicated by denoting the error component as $\text{var}(r,pr,e)$ or $\hat{\sigma}_{r,pr,e}^2$. The G coefficient for a single rater is computed as the ratio $\text{var}(p)/(\text{var}(p) + \text{var}(r,pr,e))$, and the expression for this ratio in terms of the mean squares from the one-way ANOVA is identical to that provided by Shrout and Fleiss, except for notational differences.

When scores are computed as aggregates based on judgments of multiple raters, the contribution of rater variance to observed scores is reduced. This increases the relative contribution of target variance to observed score variance, which means an increase in dependability of measurement. This is represented in the G coefficient for multiple raters by a multiplier ($1/n'_r$) for variance components involving raters, where n'_r is the number of raters for each target in the D study. This procedure can be used to forecast the dependability of ratings for any value of n'_r . Table 10.2 uses the value of $n'_r = k$ (i.e., the number of raters in the D study will be the same as that for the G study), and shows how this ratio using mean squares is again identical to that provided by Shrout and Fleiss (1979), except for notational differences.

G coefficients¹ for Shrout and Fleiss's (1979) Cases 2 and 3 are based on a crossed rating design. In this design, rater variance ($\text{var}(r)$, or $\hat{\sigma}_r^2$) can be estimated separate from the target \times rater interaction (which is confounded with random error: $\text{var}(pr,e)$ or $\hat{\sigma}_{pr,e}^2$). A GT perspective sheds some light on the choice between Cases 2 and 3, for studies with the crossed design. As Shrout and Fleiss pointed out, the choice between Cases 2 and 3 focuses on whether rater variance counts as error. The two main uses of Case 2 coefficients (which include $\text{var}(r)$ in the denominator of the G coefficient) are (a) when decisions for the study will be based on absolute scores (such as comparing observed ratings to a pre-established cutoff score) and (b) when raters are crossed with targets in the G study, but the investigator wishes to provide an estimated G coefficient for a future (D) study using a nested rating design. Thus, when investigators are mainly interested in providing an estimate of dependability for the current study (as in Contexts A and B from Table 10.1), and when raters are crossed with targets (i.e., all raters provide scores for each target), it is usually appropriate to exclude $\text{var}(r)$ from the denominator of the G coefficient (or to report $\text{ICC}(3,k)$), because $\text{var}(r)$ does not contribute to variability of observed scores in a crossed rating design.

Table 10.3 Recommendations for Reporting ICCs and G Coefficients.

<i>ICCs</i>	<i>G Coefficients</i>
<ol style="list-style-type: none"> 1. Note whether the rating design used to compute the ICC was nested (each target is judged by a different set of raters), crossed (all targets are judged by the same set of raters) or mixed. 2. If using Shrout and Fleiss's (1979) formulas (see Table 11.2), be sure to use the correct ANOVA model — one-way ANOVA for nested designs, two-way for crossed. SPSS RELIABILITY can also be used for crossed designs (Case 3). 3. If raters are crossed with targets, determine whether Case 2 (rater variance included in error term) or Case 3 (rater variance not included) is appropriate. 4. Report the ICC that reflects the level of aggregation for the scores you will be analyzing. For example, ICC(3,1) if scores are based on judgments of a single rater for each target, or ICC(3,k) if scores are composites based on k raters. 	<ol style="list-style-type: none"> 1. Determine which facets of measurement contribute to error variance, based on the intended application of scores. For many rating scales, scores are sought that generalize over raters, items, and observation occasions. 2. In the G study, vary each of these facets (e.g., each target could be judged by multiple raters, using a set of multiple relevant items, on several occasions). 3. Tabulate the results of the variance partitioning, so that readers will have access to variance estimates for facet main effects and interactions. 4. Report a G coefficient appropriate for the rating design used in the D study (actual data to be analyzed); this may be a different design than that used in the G study. Specify which sources (facet main effects and interactions) contribute to the total variance estimate for this coefficient, and the level of aggregation (e.g., number of raters per target) for each facet. 5. If the rating measure is being created for use in future substantive (D) studies, consider tabulating estimated G coefficients for possible future rating designs (e.g., crossed versus nested, and different levels of aggregation for the facets), to inform future users about the likely effects of these choices on score dependability.

For investigators who are developing a new rating measure for use in future (D) studies (Case C in Table 10.1), it may be helpful to report both the Case 2 and Case 3 coefficients (either G coefficients or ICCs), so that future users of the scale will have an indication of what the consequences will be for dependability of measurement if a nested, rather than a crossed rating design is used. When raters are nested within targets, $\text{var}(r)$ contributes to error variance, which generally results in some decrease in dependability of measurement. The difference between the Case 2 and Case 3 coefficients shows the likely magnitude of this decrease.

Summary. Because there are many types of ICCs (see Table 10.2), investigators using this method for reporting interrater reliability need to be explicit about which coefficient they are reporting and why. Users of rating measures should be aware that other sources of variance (e.g., items, occasions) contribute to error of measurement in many applications. Thus, they might opt to report a G coefficient from a G study including several facets of measurement that may contribute to measurement error, to give a clearer indication of the impact of measurement error on the findings to be reported. For investigators developing new rating measures, this approach might be particularly attractive, because it allows for tabulation of predicted G coefficients under a number of different future measurement designs. Table 10.3 provides a summary of recommendations for studies reporting either ICCs or G coefficients as indices of interrater reliability.

8. Interpretation of Findings (Contexts A, B, and C)

Methodologists have offered rules of thumb for describing the degree of agreement or reliability, based on the magnitude of the obtained coefficients. Fleiss (1981, p. 218), following Landis and Koch (1977, p. 165) recommended that kappa coefficients in the ranges .75 and higher, .40 to .75, and

below .40 be characterized as evidence of excellent, fair-to-good, and poor agreement, respectively. Typically, reliability coefficients of .80 and above have been considered to reflect good dependability of scores, with coefficients between .70 and .80 reflecting marginal dependability. Recall, however, that because multiple sources (e.g., raters, items, observation occasions) contribute to error in ratings, conventional reliability coefficients overestimate the dependability of ratings (see Desideratum 7). The rules of thumb just discussed can be useful in encouraging consistency in the labels applied to coefficients by different investigators, but ultimately the importance of these coefficients lies in what they reveal about the effects of error of measurement on the effect sizes obtained in the study.

The effects of error of measurement on study findings are complex. In general, however, when predictor and criterion variables are measured with less than perfect reliability, effect sizes involving these variables are attenuated (i.e., are smaller than they would have been under hypothetical error-free conditions). The smaller the dependability coefficient, the greater is the expected degree of effect size attenuation. Thus, interpretation of observed effect sizes should take account of the reliability of the measures that contributed to those effect sizes. This may be particularly important for comparison with effect sizes using different measurement procedures. For example, a psychotherapy process researcher may compare working alliance scores during the third session of brief psychotherapy with outcome measured at the end of treatment. If working alliance is measured via client reports, and outcome via judgments of trained interviewers, then error of measurement in each measure acts to attenuate the observed effect size. Suppose that this effect size is compared with that of an earlier study that used client report measures for both constructs and found to be smaller. It is important to ask whether this represents a substantive difference in findings, or whether it may be explained by the fact that measurement errors for the two constructs are uncorrelated in the present study, but correlated (i.e., method covariance between the two sets of client reports) in the earlier investigation. Thus, the present study may be viewed as stronger evidence for the hypothesis (despite its smaller effect size) because method covariance can be ruled out as a plausible alternative explanation for the observed correlation.

One option for quantifying the impact of measurement error is to publish effect sizes corrected for unreliability of measurement. As noted by Schmidt and Hunter (1996), it is important for investigators who use this correction to be sure that it is based on a dependability coefficient that quantifies the proportion of replicable variance taking all relevant sources of error into account (see Desiderata 7 and 10).

9. Recommendations for Future Users (Context C)

For investigators who are creating a new rating measure, it is important to provide guidance for future researchers about how to use the scale. Unlike self-report questionnaires, which are relatively standardized, rating measures offer multiple options for future users regarding the number of raters who will provide judgments of each target, the assignment of raters to targets (i.e., nested versus crossed design), and the procedures used to train raters. As noted above (see Desideratum 7), for continuous scales, a generalizability approach (reporting variance estimates and G coefficients for multiple possible rating designs) offers investigators maximum flexibility to provide guidance for future users.

10. Limitations and Suggestions (Contexts A, B, and C)

Finally, when reliability or agreement is weak or marginal, this fact should be noted as a limitation of the study in the Discussion section. Low dependability of measurement indicates that rather different scores would probably have been obtained under different rating conditions (in particular, in the case of interrater reliability or agreement, if a different set of raters had been used). As noted in

Desideratum 8, the likely consequence of unreliability in a continuous predictor or criterion variable is attenuation in the obtained effect size, so it would be expected that a replication with more reliable measures would show a stronger association. However, the impact of measurement error in other (e.g., statistical control) variables in the equation is not straightforward to predict, so that weak measurement procedures generally reduce confidence in the obtained results.

For continuous measures, investigators who report interrater reliability coefficients should acknowledge that these do not reflect the contribution of other sources of error that are relevant to evaluating dependability of measurement (see discussion of GT in Desideratum 7). Unless a generalizability analysis is undertaken, it is generally not possible to be very precise about the contribution of multiple sources of measurement error to distortion of study findings.

Note

- As noted by Hendrickson and Yin (Chapter 9, this volume), the Case 2 coefficient is technically not a generalizability coefficient (nor is it a proper intraclass correlation coefficient), because the denominator is larger than the observed score variance in a crossed rating design. Cronbach et al. (1972) referred to this denominator as an estimate of *absolute* error, whereas the denominator for Case 3 estimates *relative* error. They considered coefficients with relative error in the denominator (Case 3) to be proper G coefficients, but referred to coefficients computed on absolute error (such as Case 2 here) by the more generic name *dependability coefficients*. In Table 10.2 and the associated text, I ignore this technical issue, and refer to coefficients in all three cases as G coefficients and (following Shrout & Fleiss, 1979) as ICCs.

References

- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer Verlag.
- Campbell, J. L., Quincy, C. Q., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, 42, 294–320.
- Cicchetti, D. V., Showalter, D., & Rosenheck, R. (1997). A new method for assessing interexaminer agreement when multiple ratings are made on a single subject: Applications to the assessment of neuropsychiatric symptomatology. *Psychiatry Research*, 72, 51–63.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Cronbach, L. J., Gleser, G. C., Nanda, A. N., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 105–146). New York: Macmillan.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378–382.
- Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.
- Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, 5, 64–86.
- Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, 4, 403–424.
- Hoyt, W. T., & Melby, J. N. (1999). Dependability of measurement in counseling psychology: An introduction to generalizability theory. *The Counseling Psychologist*, 27, 325–352.
- James, R. G., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, 69, 85–98.
- Lakes, K. D., & Hoyt, W. T. (2009). Applications of generalizability theory to clinical child psychology research. *Journal of Clinical Child and Adolescent Psychology*, 38, 144–165.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- Losada, J. L., & Manolov, R. (2015). The process of basic training, applied training, maintaining the performance of an observer. *Quality and Quantity*, 49, 339–347.
- Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only a single stimulus is rated. *Journal of Applied Psychology*, 74, 368–370.
- Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, 1, 199–223.
- Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual difference constructs. *Psychological Methods*, 8, 206–224.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.
- Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22, 358–376.

11

Item Response Theory and Rasch Modeling

R. J. De Ayala

Item response theory (IRT) is a psychometric modeling paradigm used for the measurement of psychological constructs. In its simplest form an IRT model contains a single continuous latent variable that represents the construct of interest (e.g., mathematics proficiency, depression, social anxiety), which in turn is believed to determine a person's responses to a series of binary and/or polytomous questions. Both items and people are located on this continuous latent continuum. In general, a person's location on this latent continuum is estimated as a function of his or her responses to those questions and the items' location on the latent variable. From this basic model one can generalize to multiple latent construct variables that may be continuous or categorical and/or multiple item characterizations.

A typical IRT application produces an estimate of a person's location on the latent construct's continuum. However, IRT may also be used for other purposes, such as the design of an instrument with specific properties. For example, an instrument can be designed to provide highly accurate person proficiency estimates across a particular range of the latent continuum. IRT may also be used for creating an item bank for use with computerized adaptive testing or to facilitate creating alternate forms of an instrument. This latter use points toward another purpose of IRT, the equating of alternate forms of an instrument. In the desiderata and explications that follow, I distinguish between item-focused studies (e.g., item bank construction) and person-focused studies (e.g., diagnostic testing).

Although IRT typically requires larger sample sizes than those used for classical test theory (CTT) implementations, given satisfactory model-data fit IRT offers a number of advantages over the traditional CTT approach to measurement. For instance, the IRT person location estimate is not dependent on the specific instrument used for person measurement, the IRT item characterization(s) are independent of the sample of respondents, and the IRT model may be used to predict response behavior. Moreover, unlike CTT's global measure of observed score accuracy (i.e., the standard error of measurement), with IRT one knows the accuracy with which each person's location is estimated.

Because IRT models are nonlinear, obtaining estimates of the latent person and item parameters involves numerically intensive (and typically iterative) algorithms. Various software packages, such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003), MULTILOG (Thissen, Chen, &

Table 11.1 Desiderata for Item Response Theory.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. The construct of interest is defined.	I
2. IRT is justified as the appropriate measurement approach (e.g., continuous latent variable vs. categorical latent variable).	I
3. The specific model(s), with description and justification, are provided.	I, M
4. The response data are fully described, including sampling, sample size(s), demographics, and testing environment (if appropriate).	M
5. All instruments are fully described, including length, response format, and validity evidence (if appropriate).	M
6. Software and estimation approach(es) are fully specified.	M
7. Estimation problems are documented, as are details as to how they were addressed.	R, D
8. A complete description is provided of how missing data were addressed.	R, D
<i>Item-focused studies</i> (linking, item bank construction, instrument construction).	
9. Details regarding model fit analysis are provided, including those related to dimensionality, fit statistics, invariance, and model selection (if appropriate).	M, R, D
10. Details regarding item fit analysis are provided, including those related to conditional independence, functional form, fit statistics, invariance, predicted vs. observed item response functions, and handling of misfitting items.	R, D
11. Instrument calibration results are presented (item parameter estimates and/or summary statistics, total information function).	R, D
<i>Person-focused studies</i> (CAT, diagnosis, equating, vertical scaling).	
12. Person fit analysis results are presented, including fit statistics and appropriateness measurement.	R, D
13. Person location estimate results are described, including relevant standard errors.	R, D
14. Methods of equating scores on different metrics are described in detail.	M, R

* I = Introduction, M = Methods, R = Results, D = Discussion.

Bock, 2003), NOHARM 4 (Fraser & McDonald, 2012), PARSCALE (Muraki & Bock, 2003), and WINSTEPS (Linacre, 2001) are available to provide parameter estimates. Technical treatments of IRT are provided by Baker and Kim (2004), Lord and Novick (1968), and van der Linden and Hambleton (1997). Readable introductions may be found in Hambleton, Swaminathan, and Rogers (1991), as well as De Ayala (2009). The desiderata for IRT applications are presented in Table 11.1 and are explicated in the remainder of this chapter.

1. Defining the Construct of Interest

In any application of IRT one needs to define the construct(s) of interest. Because IRT is a latent variable modeling approach the authors should make clear to the reader why they believe that one or more latent variables underlie the observed behavior (i.e., in the form of item responses). In some statistical contexts (e.g., exit polling) it is unnecessary to posit the existence of a latent variable; in other cases, however, convention or theory dictates that one or more latent variables are the most meaningful conceptualization of the research questions at hand. In these cases, the operational definition of the construct should be clearly specified. Stated in other words, the linkage between, for example, a latent variable and its observed manifestations needs to be explicated.

2. Appropriateness of IRT

Although a latent variable may be invoked to explain individuals' responses to binary and/or polytomous items, this latent variable may be conceptualized as categorical, continuous, or even some combination of the two. In the case of a categorical conceptualization of the latent variable, then the use of latent class analysis (see Chapter 12, this volume) is the recommended psychometric technique. However, if the latent variable is conceptualized as continuous, then an IRT model is warranted. If the latent variable is believed to have categorical and continuous facets, then a mixture IRT model, such as the mixed Rasch model, may be used. Authors must make clear, by virtue of the hypothesized nature of the construct(s) of interest, that IRT is indeed the appropriate approach.

3. Specifying the Model(s)

After presenting the study's context, theory, purpose, and justification of why IRT is an appropriate technique, the researcher should present the IRT model(s) that will be used and a description of the model parameters. Because in some cases there may be multiple IRT models that may be applicable, the researcher should justify the models selected and articulate the implication of the selection. To elucidate this statement I will first present a taxonomy of IRT models followed by a brief discussion of the implication(s) of selecting particular models over other models.

IRT models may be classified in multiple ways. One taxonomy uses the type of response to classify the models into two broad categories, those designed solely for dichotomous (binary) response data and those for polytomous response data (e.g., from Likert response scales or rater judgments). A second classification approach reflects differences in intent. Specifically, is the researcher's intent simply to model the data or to gauge if it is possible to construct an instrument in an attempt to measure the construct of interest? This latter purpose is associated with the Rasch family of IRT models, Guttman scalogram, and Coombs unfolding. In Table 11.2 these two classification approaches are used to cross-classify several commonly used IRT models. Specifically, the IRT models are classified into whether the researcher's intent is to describe or model the data as opposed to using the model to construct the instrument and whether the model is or is not restricted to dichotomous responses; note this is a non-exhaustive list of IRT models. This grid should not be interpreted as consisting of impermeable cells. For instance, all polytomous models simplify to one of the dichotomous models and all the models listed in the first row may be seen as special cases of the models in the second row. Nevertheless, the taxonomy is a useful organizational scheme and presents two questions that need to be answered in an IRT application:

1. Is the researcher primarily concerned with modeling the data (i.e., the second row) or in using the model to determine whether or not it is possible to construct an instrument to measure the latent variable (i.e., the first row)?
2. What type or types of response data will the researcher be working with?

If the researcher adopts the perspective that the model determines whether it is possible to measure the construct (i.e., the first row), then there are a number of implications. Specifically, the models in the first row require that all items on an instrument to be approximately equally good at discriminating among respondents located at different points along the latent variable continuum. As a result, these models characterize an item only in terms of its location or, in terms of polytomous data, location(s) on the latent continuum. Therefore, when applying a Rasch model to empirical data one *might* have greater difficulty in obtaining model-data fit for some of the items because these items discriminate differently than other items and/or because of chance success on the item.

Table 11.2 Example IRT models.

<i>Intent</i>	<i>Response Type</i>	
	<i>Dichotomous</i>	<i>Polytomous</i>
Construct	Rasch/One-parameter model	Rating Scale model
	Linear Logistic Test Model (LLTM)	Partial Credit model
	Mixed Rasch model	
Describe Data	Two-parameter model	Generalized Rating Scale model
	Three-parameter model	Generalized Partial Credit model
	Multidimensional two-parameter model	Graded Response model
	Multidimensional three-parameter model	Nominal Response model Multiple-choice model

Another implication of using a model in the Rasch family involves parameter estimation. In general, with Rasch family models it is possible to obtain reasonably good item parameter estimates with smaller sample sizes than would be needed with non-Rasch models. Moreover, because there is less difficulty in estimating the item's location(s) than the other item parameter(s) (which are discussed below), convergence problems are less likely to occur with Rasch family models.

A third implication of using a Rasch family model is that results are comparatively easy to present and explain to a lay audience. For instance, two individuals with the same observed score (i.e., the sum of item responses on an instrument) will obtain the same location estimate on the latent variable. In contrast, with non-Rasch models these two individuals may receive different location estimates depending on their respective response patterns. Stated another way, with the Rasch model family all items contribute the same towards a person's location estimate (i.e., the items are equally weighted). If one wants an item to contribute, for example, twice as much as another item, then one has to explicitly assign that weight. In contrast, with non-Rasch models items are weighted by their respective discriminatory capacity.

Some individuals refer to the Rasch model as the one-parameter logistic (1PL) model, whereas others believe that because the Rasch model represents a philosophical approach to measurement not embodied in the 1PL model that the terms should not be interchanged. That is, from a Rasch philosophical perspective the Rasch model serves as the standard by which one can create an instrument for measuring a variable and, as such, is used to construct the variable of interest. This is analogous to wanting to measure how long an object is and defining a standard (e.g., a meter) with which to measure the object as well as all other appropriate objects.

The non-Rasch models in the second row may be viewed as focusing on describing the response data. To accomplish this objective these models contain one or more discrimination parameters and, in the case of the three-parameter models, an additional parameter representing the chance (so-called *guessing*) success on an item is called the *pseudo-guessing* or *pseudo-chance* parameter. These additional item parameters allow greater flexibility in modeling the data. One commonly seen multi-item parameter model is the three-parameter model (e.g., the three-parameter logistic [3PL] model). In the 3PL model each item is characterized by a discrimination parameter (α), a location parameter (δ), and a pseudo-chance parameter (χ). If one constrains χ to be zero, then the 3PL model simplifies to another common model, the two-parameter model (e.g., the two-parameter logistic [2PL] model). As can be seen from Table 11.2 these models may be extended to have multiple latent variables (e.g., the multidimensional two-parameter model).

In the above, we have assumed that we have a single population. However, we might encounter situations that involve a mixture of latent subpopulations such that there are qualitative differences

between the subgroups but within which a continuous variable exists. This situation may be modeled using a mixed Rasch model (cf., Rost, 1990, 1991; aka mixture IRT model).

The mixed Rasch model (and its variants) reflect an integration of IRT and latent class analysis. In this approach individuals are characterized by a location parameter and a latent class membership parameter. Each item has class-specific parameters (e.g., an item location within each latent class) and these parameters are continuous within each class. Our IRT model holds within each latent class, but an item's parameters may differ across latent classes. When we apply our model we can determine not only a location for each of the individual on the construct continuum within each latent class, but also to which latent class he or she is most likely to belong. As an example, assume that our latent class structure consists of one class of opportunistic cheaters and another class of non-cheaters. Our individual's location parameters characterize how much of the construct, say mathematics ability, the individual possesses given their membership in each latent class (e.g., the "opportunistic cheater" class).

At a basic level the above models may be conceptualized from a multilevel framework. That is, the models consist of two levels with item responses nested within persons (Kamata, 2001; also see De Boeck & Wilson, 2004). In the simplest framework we have a person-level model (i.e., level-2) in which we predict the parameters in the item-level model (i.e., level-1). Because an item's location is fixed across respondents there is no random effect associated with the item effect in our person-level model. At level-1 the *log odds* (or the *logit*) of a correct response on an item is predicted using $L - 1$ dummy coded predictors that function to identify an item and the coefficients constitute the item locations with respect to the item effect captured by the intercept (i.e., the L th item); L is the number of items on the instrument. This framework may be extended to include covariates as appropriate as well as by having persons nested within an additional variable, such as gender, to produce a three-level situation (i.e., items within persons within gender).

In an application's write-up, the researcher should always present the model and define its parameters rather than relying on the model's name because some models have multiple names. For example, some individuals refer to the *generalized partial credit* model as the *two-parameter partial credit* model. Moreover, presenting the model makes it clear whether a probit or a logit link function is being used. Although the difference in link functions does not affect model-data fit, which link function is used and whether any rescaling is done (e.g., the use of the D scaling constant) affects the latent continuum's metric.

4. Describing the Response Data

In the Methods section, the researcher should specify how the respondents were selected (i.e., random sampling, convenience sampling, matrix sampling, etc.) along with the sample's demographics and size. Regarding the latter, although there are a number of sample size guidelines (e.g., 100, 500, 1000, depending on the model, estimation method, and assumption tenability), these should not be interpreted as hard and fast rules. This is due, in part, to the number of the factors that need to be considered and are at times data specific; these include a researcher's model-data misfit tolerance (i.e., acceptable level of risk of failing to reject an inappropriate model), ancillary technique sample size requirements (e.g., principal component analysis), the amount of missing data, the model, the application context, the use of prior distributions for estimation as well as the estimation algorithm. An additional factor may be the number of items being calibrated. For instance, the joint maximum likelihood estimation procedure provides more accurate estimates when used with more than 25 items than with fewer items. As may be surmised, from this example some of these factors interact with one another. For instance, the greater the amount of missing data the larger the overall sample size needs to be in order to

compensate for the missing data. In general, it behooves the researcher to provide a rationale for the sample size used.

Depending on the study's context, the environment in which the instrument is administered should also be described. This includes any time constraints, administration medium (paper-and-pencil, computerized, etc.), and whether the administration was to an individual or a group.

5. Describing the Instrument(s)

Each instrument's purpose, the number of items, as well as any available validity information should be presented. Because IRT models do not require a particular item response format (e.g., multiple-choice, open-ended, ratings) it is customary to specify how the response data arose. For instance, if responses to open-ended questions are coded, a description of the rubric and an example of its application should be provided. In addition, if multiple raters are using the rubric to score the same item, then the researcher should provide a description of the rater training process, an assessment of interrater reliability, and how rating contradictions are resolved. If a Likert response scale is used, then the category labels should be specified along with whether items were reversed scored due to negative wording.

Although it is common to consider the dichotomous models as only applicable for proficiency assessment, the models are applicable for dichotomous data regardless of whether data represent correct/incorrect responses. For example, a response of 1 may be a correct response on an examination question, the successful completion of a task, a rater's judgment, or a response of TRUE on an attitude or personality item using a TRUE/FALSE response format. As such, the phrase "a response of 1" will be used herein rather than "correct response" to emphasize the generalizability of the models.

With some instruments, multiple IRT models may be used to estimate the parameters (e.g., a dichotomous model for some items and the partial credit model for other items). In these cases the researcher should specify which items are calibrated by each model. Similarly, if an instrument contains subscales, then how the multiple subscales are treated should be presented. As an example, if a unidimensional model is used with an instrument containing multiple scales, then in general each subscale should be separately calibrated. However, if the subscales are highly interrelated, then it might not be appropriate to treat each subscale individually. The researcher should make clear the approach that was used and why.

6. Estimation Approach(es)

Several estimation algorithms are available to estimate item and person parameters. The researcher should specify the estimation approach for items and, if necessary, for respondents. Furthermore, the researcher needs to indicate whether the estimation used the program's default approach or whether the defaults were changed and, if so, what the changes were.

Some of the estimation approaches make assumptions of the data, whereas others do not make these assumptions but work best in certain situations. Moreover, because certain estimation programs allow the user to select from various estimation approaches one cannot simply rely on the estimation program's name to convey information about the estimation algorithm used. Therefore, both the program name and the estimation technique used need to be specified. For instance, three commonly used approaches are *joint maximum likelihood estimation* (JMLE), *conditional maximum likelihood estimation* (CMLE), and *marginal maximum likelihood estimation* (MMLE). These latter two approaches are available in the estimation program OPLM (Verhelst, Glas, & Verstralen, 1995). A fourth estimation strategy uses unweighted least squares to fit a polynomial that approximates

a two-parameter normal ogive model (i.e., a two-parameter model using a probit link function). This strategy is typically used for estimating item parameters for a dichotomous multidimensional model, although it may also be used with dichotomous unidimensional models.

Although IRT models do not make any distribution assumptions, MMLE makes an assumption about the respondent distribution. Typically, this assumption is that the respondents come from a normally distributed population with respect to the latent variable. In contrast, JMLE does not make assumptions about the respondents' distribution. It should be noted, however, that JMLE has known inherent weaknesses that are exacerbated under certain conditions (e.g., short instruments and small samples) and, as a consequence, may encounter estimation difficulties with certain data sets and certain IRT models.

Some of the estimation strategies, including MMLE and unweighted least squares, only provide item parameter estimates. Therefore, if one needs estimates of the respondents' locations it would be necessary to perform a second step. Some of the commonly available person estimation approaches are *maximum likelihood estimation* (MLE) and Bayesian estimation such as *maximum a posteriori* (MAP) or *expected a posteriori* (EAP). The Bayesian approaches make a respondent distribution assumption and, as a result, the degree of regression of the person location estimate toward the mean will, in part, be dependent on the (prior) distribution's parameters. However, unlike using MLE for estimating an individual's location the Bayesian approaches will provide a person location estimate for each individual. Therefore, the estimated person locations will differ across the estimation procedures (although they will be highly linearly related with one another). Again, because some programs (e.g., BILOG-MG) provide multiple person estimation algorithms, simply specifying the program's name would be insufficient to inform the reader of which person estimation approach was used. As a result, both the estimation program and the estimation algorithm used for persons should be made clear in the Methods section.

The decision of which algorithm to use depends on a number of factors, including the model being used, the instrument length, available calibration software. For instance, CMLE is only an option with the Rasch family of models. With other models one could use JMLE, MMLE, or unweighted least squares. However, if one's instrument is 10 items long, for example, then JMLE would not be the best approach to use because research has shown that instruments should be at least 25 items long for proper estimation with JMLE. Moreover, when performing JMLE calibrations for the two- and three-parameter models there should be at least 1000 respondents to reduce bias in the parameter estimates. For the Rasch model family this bias can be ameliorated with certain JMLE programs. For non-Rasch family models, in general, most calibrations are currently being performed using MMLE.

7. Addressing Estimation Problems

Estimation problems are more likely to occur with models that include a discrimination parameter and/or a pseudo-guessing parameter. Therefore, the following will apply primarily to non-Rasch models and the use of MMLE.

In some situations it is possible to experience difficulty in estimating an item's discrimination parameter (α). For example, for some items the estimate of α may drift off to infinity; this is sometimes referred to as an example of a *Heywood case*. There are several approaches that one might use to aid in estimating an item's discrimination. One approach uses a prior distribution (e.g., a lognormal distribution) for the estimation of α . Although, in general, the use of a prior distribution produces estimates that may be regressed toward the prior distribution's mean, the use of a prior distribution with discrimination parameter estimation has a less serious impact than in the case of person and item location parameters. Unless otherwise specified by the user, BILOG imposes

a lognormal prior distribution when estimating α for the two- and three-parameter models. An alternative strategy is to impose an upper limit on the values that the estimated discrimination parameters may take on. This is the approach used in some JMLE programs.

Difficulty in estimating an item's lower asymptote or pseudo-guessing parameter (χ) may occur because of (1) problems in estimating an item's other parameters, (2) because of the other parameters' estimates (e.g., a low estimated α parameter), (3) because the item is located at the lower end of the continuum (i.e., a very easy test item), and/or (4) because there are insufficient data at the lower end of the continuum with which to estimate an item's χ . In this latter case, there may be several different combinations of an item's parameters that produce item response functions (IRFs) whose lower asymptotes are similar to one another even though the item parameter estimates may be vastly different from one another.

Different approaches for handling these estimation difficulties involve using a prior distribution or fixing χ to a specific value. In this latter case, this constant (common) value for χ may be set arbitrarily (e.g., $1/m - 0.05$, where m is the number of item options), by averaging the non-problematic χ estimates, by averaging the χ estimates for items located at the lower end of the continuum (i.e., the easy items), or by fixing the lower asymptote to some nonzero value determined by inspecting the lower asymptote of empirical item response function. In addition, a "stability" criterion may be invoked to determine whether an item's χ parameter should be estimated at all (or assigned a constant value). That is, χ is estimated only when $\delta - 2/\alpha > -2.5$, where δ is the item location.

As mentioned above, we may also use a prior distribution for estimating items' χ parameters. The use of a prior distribution for estimation of χ can lead to reasonable parameter estimates for the model. Moreover, the regression toward the mean phenomenon that typically occurs when using a prior distribution is not as problematic in estimating χ as it is when estimating person and item location parameters. The use of a prior distribution is recommended as the first strategy to facilitate the estimation of χ . By default BILOG uses a beta distribution as a prior for estimating the IRF's lower asymptote, χ , for the three-parameter model.

With polytomous data it is possible that one or more of an item's response categories may not be attractive and may never be chosen by respondents. These are sometimes referred to as *null categories*. In general, it is not possible to estimate the parameters for a category that does not have any observations. However, some software packages may provide "estimates" for the null category. If a null category occurs, then one should ignore the null category's parameter estimates and recalibrate the item set specifying the appropriate number of categories actually observed for each item.

The preceding has been concerned with item parameter estimation problems. However, it is possible to experience problems when estimating person locations. If this occurs it is typically associated with using MLE. Although it is possible with MLE to obtain nonfinite person location estimates with poorly behaved likelihood functions, the most common problem to occur with binary response strings is an infinite person location estimate due to either responses of all 1s or responses of all 0s. There are several strategies that may be used to perform MLE with response strings consisting of all 1s (i.e., perfect response strings) or all 0s (i.e., zero response strings). The gist of these strategies is to modify the response strings to introduce some nonuniformity. One approach, the *half-item rule*, assigns 0.5 to the item with the smallest location value for a uniformly 0 response vector and to the item with the largest location value and for a uniformly 1 response vector. For example, assuming five items in increasing order of their locations, then a zero response string would become [0.5 0 0 0 0] and a perfect response string would be [1 1 1 1 0.5]. The strategy used for addressing perfect response or zero response strings should be specified in the write-up of the study. In contrast and as mentioned above, with EAP and MAP, person location estimates are available for all response vectors.

8. Missing Data

IRT is concerned with modeling *observed* responses. However, in working with empirical data one will, at times, encounter situations where some items do not have responses from all individuals in the calibration sample. Some of these missing data situations may be considered to be missing by *design* or *structurally missing*. For example, one may administer an instrument to one group of people and an alternate form of the instrument to another group. If these two forms have some items in common, then the calibration sample can consist of both groups. As a result, the data contain individuals who have not responded to all the items on both forms. In situations where the nonresponses are missing by design, these missing data may be ignored because of the IRT properties of person and item parameter invariance. However, when nonresponses are not structurally missing, then one needs to consider how to treat these nonresponses.

In general, missing data (e.g., omitted responses) may be classified in terms of the mechanism that generated the missing values: *missing completely at random* (MCAR), *missing at random* (MAR), and *nonignorable*. MCAR refers to data in which the missing values are statistically independent to the values that could have been observed as well as to other variables. In contrast, when data are MAR then the missing values are conditionally independent on one or more variable(s). Nonignorable missing values are data for which the probability of omission is related to what the response would be if the person had responded.

In the IRT context there are various reasons why an individual's response vector might not contain responses to each item. For instance, and as mentioned above, in the missing by design case one has *not-presented* items. This missingness due to not-presented items arises in, for example, adaptive testing or the simultaneous calibration of multiple forms. These nonresponses represent conditions in which the missingness may be ignored for purposes of person location estimation. Therefore, the estimation is based only on the observed responses.

A second situation that will produce missing data occurs when an individual has insufficient time to answer the item(s). These *not-reached* items are (typically) identified as collectively occurring at the end of an instrument (this assumes the individual responds to the test items in a serial fashion) and represent *speededness*. Although IRT should be applied to unspeeded tests/instruments, if we knew which items the examinee did not have time to consider, then these not-reached items could be ignored for person location estimation because they contain no readily quantifiable information about the individual's location on the latent continuum (Lord, 1980). Therefore, when one has (some) missing data due to not-reached items, then the person's location can be estimated using only the observed responses. This should not be interpreted as indicating that IRT should be applied to speeded instruments, nor that the item parameter estimates for the not-reached item(s) are unaffected by being speeded. In fact, speeded situations may lead to violation of the unidimensionality assumption and, subsequently, biased item parameter estimates. For example, research has shown that the speeded items' α and δ parameters are overestimated and the χ parameters are underestimated. Because of the α overestimation the corresponding item information, and therefore the instrument's total information, are inflated. Identifying the speeded items as not-reached within BILOG mitigates the bias in item parameter estimation.

The third situation that will produce missing data occurs when an examinee intentionally chooses not to respond to an item for which he or she does not have an answer. These *omitted responses* represent nonignorable missing data. Again, assuming that an individual responds in a serial fashion to an instrument, omitted responses may be distinguished from not-reached items by appearing throughout the response vector and not just at its end. In the context of cognitive assessment, omitted responses may not be ignored because individuals could obtain, for example, as high a proficiency estimate as they wished by simply answering only those items they had confidence in correctly answering.

Research has shown that, with dichotomous data, omits should not be treated as incorrect nor should they be ignored. In these cases using a fractional value of 0.5 in place of omitted values leads to improved person location estimation compared to treating the omits as responses of 0 or using some other fractional value. By using this fractional value one is simply imputing a “response” for a binomial variable and thereby attempting to “smooth irregularities” in the likelihood function. Additionally, SAS’s PROC MI or SPSS’s MISSING VALUE ANALYSIS-EM may be used to impute values for the omitted responses and the resulting complete data calibrated.

The usefulness and accuracy of imputation approaches depends on the flexibility of the practitioner’s calibration software. That is, these approaches will yield decimal values and with some calibration software these values will need to be converted to integers. An alternative approach is to perform hot-decking. That is, for each case with missing data another case is found that is similar in characteristics to the case with the missing value(s), but has responses for the item(s) in question. The responses from this second case are substituted for those in the case with missing values. If one has multiple matching cases, then one selects the matching case at random. A third strategy that may be fruitful in some situations is to treat the omission as its own response category and apply a polytomous model such as the multiple choice or the nominal response models.

There are some issues in the treatment of omits that the practitioner should be aware of. For instance, in the context of proficiency assessment all the imputation procedures that produce “complete” data for analysis are, in effect, potentially giving credit (partial or otherwise) for an omitted response. With the substitution of a fractional value for the omitted response (e.g., 0.5) a second issue is the use of the same imputed value for all omits assumes that individuals located at different points on the latent variable continuum can all be treated the same. These issues are mentioned so that the practitioner understands the assumptions that are being made with some of the missing data approaches discussed above. However, they may or may not be of concern to an individual practitioner. Moreover, when IRT is used in personality testing or with attitude or interest inventories these may be less critical due to the assessment’s lower stakes. A third issue that should also be noted is that omits tend to be associated with personality characteristics and demographic variables as well as proficiency level. As such, in those situations where information on these variables is available one might wish to use this information as covariate(s) in the imputation process; use of these covariate(s) might or might not have any meaningful impact on the person location estimates.

It is good practice when calibrating a data set to identify items without responses by some code. For instance, in the data file not-reached items may be identified by a code of, say, 9, not-presented items by a code of 8, omitted items by a code of 7. With certain calibration programs (e.g., BILOG 3, BILOG-MG, MULTILOG) any ASCII character may be used (e.g., the letters ‘R’ for not-reached, ‘P’ for not-presented, and ‘O’ for omit). For BILOG omitted responses must be identified as such, whereas with other programs (e.g., MULTILOG) any response code encountered in the data file that was not identified as a valid response is considered to reflect an omitted item.

In practice, the user should report the strategy used to address nonignorable omitted responses as well as the rationale for why it was selected. As an aid in deciding on a strategy the practitioner may want to perform the estimation with and without the missing data strategy and compare the results to determine the impact of the strategy. Furthermore, this comparison (e.g., determining the impact of the strategy) may be facilitated by simulating data using the calibration model and the estimated item parameters.

9. Model Fit

Ignoring the socio-political factors that in practice may determine which model is used, model selection would, in general, involve considering the latent structure (e.g., the number of latent

variables, whether the latent variables are continuous and/or categorical), response data characteristics (e.g., dichotomous, polytomous), and the importance and necessity of modeling chance success as well as differential discrimination across items. Therefore, model selection would initially involve each of these factors as well as assessing the tenability of the corresponding assumptions (e.g., a model's dimensionality assumption). However, after selecting one or more models there remains the issue of model-data fit.

If the dimensionality analysis suggests that one or more models based on a single latent variable are appropriate, then one approach to assessing model-data fit is by using the likelihood ratio statistic, G^2 . A statistically nonsignificant G^2 provides evidence supporting that the data are consistent with what a model would predict. In some cases the sample size and the typical instrument length will yield significant G^2 values. However, in these situations convention has been to still use G^2 to compare (nested) models of varying complexity. (Calibration programs typically produce the log likelihood value that would be used in calculating G^2 .) For instance, because the one-parameter model is nested within the two-parameter model one may use the G^2 statistic to determine if the two-parameter model's varying discrimination parameter is necessary. If the G^2 is not statistically significant, then this would indicate that it is not necessary to provide each item with its own discrimination parameter and the simpler one-parameter model would be favored. Conversely, if G^2 is statistically significant, then the additional item parameter in the two-parameter model would be necessary. Similarly, the necessity of including a pseudo-guessing parameter for each item (i.e., the three-parameter model) could be assessed by comparing the model's fit with that of models without this item parameter (i.e., the one- or two-parameter models). The degrees of freedom for G^2 would be the difference in the number of item parameters between the two models.

With regard to polytomous response data, all the models listed in Table 11.2 (except for the graded response model) may be viewed as nested within the multiple-choice model. As such, the rating scale model is subsumed by the generalized rating scale, partial credit, and generalized partial credit models, the partial credit model is nested within the generalized partial credit model, the generalized partial credit model is nested within the nominal response model, and the nominal response model is nested within the multiple-choice model. Therefore, and as was the case with the dichotomous unidimensional models, the G^2 statistic can be used to compare various polytomous models; degrees of freedom associated with G^2 would be the difference in the number of item parameters between the compared models.

With multidimensional data the G^2 model comparison strategy is currently difficult to implement because the most popular software for parameter estimation does not produce a log likelihood value for the model solution. As a consequence, selecting between the multidimensional two- and three-parameter models would be based on a more heuristic approach. In general, instruments that contain items without a correct answer, such as, those found on a personality scale, should be adequately modeled using the multidimensional two-parameter model because there is no reason to believe that a respondent would guess at a response. However, for instruments that contain items on which chance success is a possibility, then one might use the *unidimensional* three-parameter model to get a sense of the degree to which chance success is evident by examining the χ parameter estimates. If this degree is very small, then using the multidimensional two-parameter model may be adequate for modeling the data; otherwise, a multidimensional three-parameter model would be called for.

In contrast to G^2 , one may use one or more information criteria to assess fit. These indices, based on the *log likelihood* (i.e., $-2\ln L$) function used in parameter estimation, are not statistical significance tests. Some common indices are the Akaike information criterion (AIC; Akaike, 1974), consistent AIC (CAIC; Bozdogan, 1987), the Bayesian information criterion (BIC; Schwarz, 1978), and the sample-size adjusted Bayesian information criterion (SABIC; Sclove, 1987). For each of

these information indices we select the model with the lowest index value among a set of competing models. (If more than one model shares the same index value, then we select the most parsimonious model.) These indices reflect relative fit and are meaningful when used for making comparisons across a set of models applied to the same data. Generally speaking, as model complexity increases fit improves and the maximum of the log likelihood function decreases reflecting better fit. For example, a two-class mixture IRT model will have a maximum log likelihood that is larger than that of a three-class mixture IRT model.

Although these statistics all utilize the $-2\ln L$ of the fitted model, they differ in the penalty imposed for model complexity (i.e., the number of parameters estimated) and, in some cases, sample size. The addition of the penalty seeks to compensate for the decreasing log likelihood associated with increasing model complexity. As such, these indices try to balance model-data fit as captured in the log likelihood function with the selection of a parsimonious model by imposing a penalty for model complexity. For AIC the penalty is based solely on the number of model parameters, whereas with CAIC, BIC, and SABIC the penalty also involves the sample size. BIC, CAIC, and SABIC differ in the implementation of the sample size penalty. With respect to BIC, as the number of parameters estimated and/or the sample size increases so does the penalty. In other words, BIC's penalty leads BIC to favor models that have fewer parameters over models that are more complex, this is particularly true with large samples. Unlike AIC, as sample size increases BIC tends to select the correct model (Haughton, 1988). As its name implies SABIC's penalty seeks to moderate the sample size effect used in BIC (or CAIC); CAIC's penalty is more severe than BIC for a given number of parameters and sample size. Because of the use of sample size and/or the number of parameters estimated, the AIC, CAIC, BIC, and SABIC indices will not always agree with one another in model selection. Thus, they should be considered as providing guidance in model selection rather than determining which model should be selected.

10. Item Fit Analysis

Although one might have evidence of model-data fit this does not necessarily mean that one has evidence of fit for each item. Stated another way, if one has item-data fit for each item, then one will have model-data fit; however, the converse does not have to be true. Whether or not one or more misfitting items adversely affect the model-data fit analysis depends on the number of problem items and the instrument's length. For instance, one may have satisfactory overall model fit, but upon further investigation finds that a couple of items exhibit differential performance across the manifest groups of males and females. Therefore, after selecting a model and/or obtaining overall model-data fit one proceeds to perform item level analyses. In addition, in some cases, model-level *misfit* may be diagnosed by examining the item-level fit statistics (e.g., chi-square, INFIT, OUTFIT).

Item level fit analyses should involve both statistical indices and graphical analyses. The statistical analyses depend, in part, on the calibration program because different programs provide different statistics or none at all. The graphical analyses allow an examination of IRT's functional form assumption as well as certain model assumptions (e.g., constant discrimination). In the following I begin with some statistical indices and then proceed to graphical methods.

Most statistical indices are variants of a chi-square statistic. Programs such as WINSTEPS produce two fit statistics, INFIT and OUTFIT, for examining fit; these statistics may also be used for model-data fit analysis. Other programs such as BILOG-MG produce a chi-square fit statistic. I will first discuss INFIT and OUTFIT.

INFIT is a weighted fit statistic based on the squared standardized residual between what is observed and what would be expected on the basis of the model. These squared standardized residuals are information weighted and then summed across observations for the j th item; the weight is

$p_j(1 - p_j)$ where p_j is the probability of a response of 1 according to, for example, the Rasch model. OUTFIT is also based on the squared standardized residual between what is observed and what would be expected, but the squared standardized residual is not weighted when summed across observations. As such, OUTFIT is an unweighted standardized fit statistic. Both statistics are averaged to produce mean-square statistics. INFIT and OUTFIT each have a range from 0 to infinity with an expectation of 1; their distributions are positively skewed.

These two statistics differ in their sensitivity to where the discrepancy between what is observed and what is expected occurs. For instance, responses by persons located near an item's location estimate that are consistent with what would be expected according to the model produce INFIT values close to 1 (given the stochastic nature of the model). However, responses by persons located near the item's location estimate that are inconsistent with what would be expected lead to large INFIT values. In short, INFIT is sensitive to unexpected responses near the item's location. In contrast, OUTFIT has a value close to its expected value of 1 when responses by persons located away from an item's location estimate are consistent with what is predicted by the model (again, given the stochastic nature of the model). However, unexpected responses by persons located away from an item's location estimate lead to OUTFIT values substantially greater than 1. Therefore, OUTFIT is sensitive to, say, a high ability person incorrectly responding to an easy item or a low ability person correctly responding to a hard item.

Although there are various interpretation guidelines, one guideline states that values from 0.5 to 1.5 are "acceptable" with values greater than 2 warranting closer inspection of the associated item. However, using a common cutoff value does not necessarily result in the fit statistic (either INFIT or OUTFIT) having correct Type I error rates. Some have suggested taking sample size (N) into account when interpreting INFIT and OUTFIT. Specifically, INFIT and OUTFIT would use $1 \pm 2/\sqrt{N}$ and $1 \pm 6/\sqrt{N}$ as cutoff values, respectively. Alternatively, given INFIT and OUTFIT's expectation and their range it is clear that there is an asymmetry in their scales. Therefore, INFIT and OUTFIT can be transformed to have a scale that is symmetric about 0.0. The result of this transformation is a standardized (0, 1) fit statistic, ZSTD. On this metric good fit is indicated by INFIT ZSTD and OUTFIT ZSTD values close to 0. Because the ZSTD values are approximate t statistics, as sample size increases they approach z statistics. As such, values of +2 are sometimes used for identifying items that warrant further inspection. See Linacre and Wright (2001) and Smith (2004) for more information on the INFIT and OUTFIT statistics and their transformations.

As mentioned above, BILOG-MG produces a chi-square fit statistic. This statistic is unreliable when calculated on instruments with fewer than 20 items (Zimowski et al., 2003). These chi-square statistics are based on combining individuals into, by default, 9 intervals on the basis of their Bayes location estimates; the user may change the number of intervals. The chi-square statistic tests the null hypothesis that an item's data are consistent with the model. In other words, item fit is associated with statistically nonsignificant chi-square values. As is generally true, failure to reject the null hypothesis does not imply that the model is correct, but only that there is insufficient evidence to reject the model.

In addition to using item fit statistics one should also compare the agreement between the empirical IRF with the model predicted IRF; for polytomous models one would compare the empirical and predicted option response functions (ORFs). This graphical examination should be viewed as a complementary approach to assessing item fit with fit statistics and not a replacement for fit statistics.

What defines agreement between the predicted and empirical IRFs is not absolute. One may use error bands to define reasonable agreement between the two IRFs. That is, if the empirical IRF falls within the error bands, then there is agreement between the empirical and the predicted IRFs (i.e., item-data fit). Because the width of these error bands is a function of the standard error one needs

to decide on the number of standard error units that would indicate agreement. For example, the error bands might reflect two standard errors above and below the predicted IRF.

It should be noted that the agreement between the empirical and predicted IRFs is a matter of degree and only informs our judgment of fit. For instance, sometimes we find that the empirical IRF reflects an ogival pattern that shows close agreement with the predicted IRF for a substantial range of the continuum, but disagreement at, for example, the lower end of continuum (say below -2). Depending on the application this lack of fit at and below -2 may not be reason for concern. In short, different situations may be more amenable to or accepting of a certain degree of less than perfect fit. Another consideration is the number of intervals used in creating the empirical IRFs. For example, with a small number of intervals we might observe strong agreement between the empirical and the predicted IRFs, but with a larger number of intervals the degree of agreement is not as strong all other things being equal. Also, in making our judgment of fit we recognize that the choice of say, two, standard errors for defining the error band width is a reasonable, but arbitrary, decision. Again, all of this information is used to inform our judgment of fit along with the context (e.g., the number of items on the instrument, the number of items exhibiting "weak agreement," the number of respondents, the amount of missing data, the purpose of the application, and so forth).

So far we have been concerned with assessing item-level fit. However, we can also use item-level information to assess the conditional independence assumption that all IRT models make. The gist of this assumption is that responses to two items are statistically independent of one another after conditioning on the latent person variable(s). Although there are a number of different statistics that may be used for determining the tenability of this assumption, one simple approach is to use the Q_3 statistic. The Q_3 statistic is the correlation between the residuals for a pair of items. The residual for an item is the difference between the individuals' observed responses and their corresponding expected responses on the item. Therefore, after fitting the model the Pearson correlation coefficient is used to examine the linear relationship between pairs of residuals. For instance, with dichotomous data on the j th item the observed response (x_{ij}) is either a 1 or 0 and the expected response is the probability (p_j) given by the IRT model. Symbolically, the residual for person i for item j is $d_{ij} = x_{ij} - p_j(\hat{\theta}_i)$ and for item k it is $d_{ik} = x_{ik} - p_k(\hat{\theta}_i)$; $\hat{\theta}$ is the person location estimate. The Q_3 statistic is the correlation between d_{ij} and d_{ik} across persons (i.e., $Q_{3jk} = r_{d_{ijk}d_{ik}}$).

If $|Q_3|$ equals 1.0, then the two items are perfectly dependent. A Q_3 of 0.0 is a necessary, but not a sufficient condition for independence because one may obtain a $Q_3 = 0$ because the items in an item pair are either independent of one another or because they exhibit a nonlinear relationship. Therefore, Q_3 is useful for identifying items that exhibit item *dependence*. Under conditional independence Q_3 should have an expected value of $-1/(L - 1)$, where L is the number of items on the instrument.

In some situations one may explain item dependence in terms of multidimensionality. That is, the dependency between two items is due to a common additional latent variable (e.g., test-wiseness in achievement testing). If two items are conditionally independent, then their interrelationship is completely explained by the latent structure of the model. However, it has been shown that if one applies a unidimensional model to bidimensional data, then items that are influenced by both latent variables will show a negative local dependence and items that are affected by only one of the two latent variables will show a positive local dependence. If only one of the latent variables is used, then the items that are influenced only by that underlying variable will show a slight negative local dependence. To obtain a large Q_3 value one needs to have similarity of items parameters for the items in question and the items need to share one or more unique dimensions. Therefore, similarity of parameters is a necessary, but not a sufficient condition, for obtaining a large Q_3 value. To summarize, if one determines that the various item pairs are exhibiting conditional independence, then one also has evidence support model-data fit.

Another facet of item-level fit analysis involves obtaining evidence of item parameter invariance. In the current context, invariance refers to one or more sets of item parameters that are interchangeable within a linear transformation. Although, theoretically, IRT item parameters are invariant, whether invariance is realized in practice is contingent on the degree of model-data fit. Therefore, obtaining evidence of invariance can be used as part of a model-data fit investigation. The quality of model-data fit may be assessed by randomly dividing the calibration sample into two subsamples. Each of these subsamples is separately calibrated and their item parameter estimates compared to determine their degree of linearity. This comparison can simply involve calculating the correlation coefficients between the subsamples' calibration results. One would have invariance evidence if the correlation coefficients are large (e.g., greater than 0.9). This approach would be applied across items for each of the items' parameters (i.e., a correlation coefficient for item discrimination across subsamples, another for item difficulty, etc.).

The correlation coefficient approach for invariance assessment has its disadvantages. One potential disadvantage is that the coefficient does not identify individual problem item(s). This issue may be addressed by obtaining the scatterplot for each coefficient. Another disadvantage is that it is not possible to simultaneously examine the interaction of the item's parameter estimates because the correlation is calculated for each item parameter. However, more sophisticated approaches for comparing an item's response functions based on each subsample's estimates allow one to simultaneously evaluate an item's multiple parameter estimates.

To simultaneously compare an item's multiple parameter estimates across subsamples one can use some of the *differential item functioning* (DIF) methods. DIF methods are used to detect items that are functioning differently across manifest groups of individuals. Typically, these manifest groups reflect majority/minority groups (e.g., males and females, African Americans and Caucasian, Hispanics and non-Hispanics). However, these manifest groups can also be two randomly created subsamples. Some DIF approaches that could be used for this purpose are the likelihood ratio method of Thissen, Steinberg, and Wainer (1988), Lord's chi-square (Lord, 1980), and the exact signed area and H statistic methods (Raju, 1988, 1990).

11. Instrument Calibration

Which calibration results one presents and the completeness of the presentation depends on the study's purpose. In general, it is good practice to describe one's instrument statistically and/or graphically. At a minimum one should provide descriptive statistics for each type of item parameter. For example, one would provide the mean, range, the standard deviation, and so forth for the item parameter estimates. Moreover, one can graphically present the instrument's total (i.e., test) information function to show where on the continuum the instrument is expected to provide the most accurate person location estimates.

When the purpose of the study involves linking different metrics (see Desideratum 14) then the researcher should provide information about the metric transformation coefficients. Similarly, in those cases where person location estimates are transformed to the total score metric then the researcher should provide the total (i.e., test) characteristic function (TCF). In some cases, the item parameter estimates for all the items can be presented in a table either in the body of the paper or in an appendix; the corresponding standard errors should also be provided.

With the Rasch family of models it is possible to present an item–person map (also known as a variable map). The item–person map shows how the distributions of respondents and items relate to one another. By comparing the item locations to the person distribution one can obtain an idea of how well the items are functioning in measuring persons' latent trait(s). In short, this graph allows one to see not only how well the respondents' distribution matches the range of the instrument,

but also provides an idea of how well the items are distributed across the continuum. Using this information one may anticipate where on the continuum one may experience greater difficulty in estimating person as well as item locations.

12. Person Fit Analysis

Analogous to item fit analysis, different calibration programs provide different person fit information. For instance, some of the calibration programs for the Rasch family of models produce the INFIT and OUTFIT statistics (Desideratum 10). In terms of person fit, these statistics are interpreted in a fashion analogous to their use with item fit. Specifically, responses on items located near the person's estimated location, $\hat{\theta}$, that are consistent with what would be expected produce INFIT values close to 1. However, responses on items located near the person's $\hat{\theta}$ that are inconsistent with what would be expected lead to large INFIT values. That is, INFIT is sensitive to unexpected responses near the person's $\hat{\theta}$. In contrast, OUTFIT has a value close to its expected value of 1 when responses on items located away from a person's $\hat{\theta}$ are consistent with what is predicted by the model. Conversely, OUTFIT values substantially greater than 1 arise because of unexpected responses on items located away from a person's $\hat{\theta}$.

Other fit statistics include, but are not limited to, the UB statistic (Smith, 1985) and Klauer and Rettig's (1990) chi-square statistic. The UB statistic may be standardized and a standard normal table can be used to provide screening values that would aid in identifying individuals that warrant further scrutiny. For example, a large UB statistic would indicate a person that is behaving inconsistent with the model (i.e., a misfitting person). The Klauer and Rettig chi-square statistic is asymptotically distributed as a chi-square. Therefore, the standard chi-square table of critical values would be used to identify a misfitting person. Klauer and Rettig have evaluated the significance of their statistic with a α level of 0.10. In general, if one or more persons are found to be misfitting, then at the very least they should be removed from the sample and the response data re-calibrated to determine their impact on the item parameter estimation. If it is determined that misfitting persons have minimal impact on the calibration results, then the researchers may choose to report the results that include the misfitting people.

A complementary graphical approach for person-fit assessment is the *person response function* (PRF). The PRF presents the relationship of the probability of a person's response pattern and the item locations. In addition to being used for identifying misfitting individuals, the PRF may be used to identify a particular item or set of items for which person-item fit is problematic as well as for providing diagnostic information, such as inattention, guessing, identifying copying, and so on (Trabin & Weiss, 1983). Typically, the PRF is assumed to be a nonincreasing function of the item locations. Departures from this monotonicity assumption are taken as indicators of person-model misfit for all or some subset of the instrument's items. To examine person fit, one compares a person's observed PRF (OPRF) with his or her expected PRF (EPRF). To obtain an individual's EPRF one uses his or her $\hat{\theta}$ and the item parameter estimates to calculate the person's probability of a response of 1. To obtain the individual's OPRF we first group the items in terms of the similarity of their locations and then determine the proportion of items in each group for which the individual has a response of 1. Large discrepancies between the individual's OPRF and EPRF reflect an individual who is behaving inconsistently with the model.

An alternative to the fit statistics' perspective (i.e., is the person behaving consistent with the model) is *appropriateness measurement*. In appropriateness measurement one asks, "What is the appropriateness of a person's estimated location as a measure of his or her true location (θ)?" For example, assume that a person has a response pattern of missing easy items and correctly answering more difficult items. One might ask, "Did this pattern arise from the person correctly guessing on

some difficult items and incorrectly responding to easier items or does this reflect a person that was able to copy the answers on some items?" Various statistically-based indices have been developed to measure the degree to which an individual's response pattern is unusual or is inconsistent with the model used for characterizing his or her performance.

One appropriateness index, l_z , has been found to perform better than other measures. This index is based on the standardization of the person log likelihood function, and allows the comparison of individuals at different θ levels on the basis of their l_z values. Although l_z is purported to have a unit normal distribution, this has not been found to be the case for instruments of different lengths. As a result, it is not advisable to use the standard normal curve for hypothesis testing with l_z . Nevertheless, various guidelines exist for using l_z for informed judgment. In general, a "good" l_z is one around 0.0. An l_z that is negative reflects a relatively unlikely response vector (i.e., inconsistent responses), whereas a positive value indicates a comparatively more likely response vector than would be expected on the basis of the model.

13. Person Location Estimates

In a classification or certification situation the decision about the individual is provided. For non-classification or non-certification contexts one typically presents the person location estimates on a transformed metric to eliminate negative person estimates and to make the scores more easily interpreted than they would be on the untransformed metric. This transformed metric may be a total score scale (e.g., a number correct scale) that ranges from 0 to the number of items on the instrument, or another metric that has been adopted (e.g., T-scores, College Board Score Scale, a proprietary scale). In other situations, particularly when not presenting the person estimate to the public, the person's location estimates may be left on the standard score-like metric that allows for both negative and positive location estimates. It is good practice to provide the estimate's standard error. How one transforms the θ standard metric to another metric is discussed next in Desideratum 14.

14. Metric Definition and/or Transformation

Because of the indeterminacy of the parameter metric calibration, programs typically use either person-centering or item-centering to identify the calibration model. The net effect of this approach is that the metric is defined relative to the sample used for the parameter estimation. Assuming acceptable model-data fit, then the administration of an instrument to two distinct samples will more than likely result in parameter estimates that are not identical because each sample defines its own metric. However, these two metrics are linearly related to one another (i.e., the metric is determined up to a linear transformation). As such, it is possible to transform one metric to another so that the interpretation of the estimates is freed from the particular sample used for estimation.

To linearly transform one metric to another involves using *metric transformation coefficients* (also known as *equating coefficients*). In general, the linear transformation from one metric to another for both person and item locations (or their estimates) is

$$\xi^* = \zeta(\xi) + \kappa, \quad (1)$$

where ζ and κ are the unit and location coefficients, respectively. The ξ term represents the parameter (or its estimate) on the untransformed or *initial metric* and ξ^* represents the same parameter transformed to the *target metric*. The target metric (sometimes called the *common metric*) is the metric onto which all other metrics are transformed. For example, ξ can represent the item location parameter, δ_j (or its estimate, $\hat{\delta}_j$), on the initial metric and ξ^* is δ_j^* (or $\hat{\delta}_j^*$) on the target metric.

To transform the initial metric's item discrimination parameter, α_j , to the target item discrimination parameter metric, α_j^* , we use

$$\alpha_j^* = \frac{\alpha_j}{\zeta}. \quad (2)$$

(The discrimination parameters' estimates may be used in lieu of α .) The IRFs' lower asymptote parameters (or their estimates) are on a common [0, 1] metric and do not need to be transformed.

In those cases where either a proprietary scale (e.g., the College Board scale) or a commonly defined scale (e.g., the T-score) is the target metric, then ζ and κ are given by the target metric's definition. For instance, for the T-score scale, $\zeta=10$ and $\kappa=50$. However, when ζ and κ are not given by the target metric then it is necessary to estimate ζ and κ . For example, this would be the case when we are linking two metrics to one another.

Multiple approaches for determining the values of ζ and κ have been developed. One approach obtains the metric transformation coefficients by using the mean and standard deviations of the common items; this approach is sometimes called *linear equating*. Specifically, the coefficient ζ is obtained by taking the ratio of the target to initial metric standard deviations (s) of the locations:

$$\zeta = \frac{s_{\delta^*}}{s_\delta}, \quad (3)$$

where s_{δ^*} is the standard deviation of the item locations (or their estimates) on the target metric and s_δ is the standard deviation of the item locations (or their estimates) on the initial metric. Once ζ is determined the other coefficient, κ , is obtained by

$$\kappa = \bar{\delta}_j^* - \zeta \bar{\delta}_j. \quad (4)$$

where $\bar{\delta}_j^*$ is the mean of the item locations on the target metric and $\bar{\delta}_j$ is the mean of the item locations on the initial metric. Once the metric transformation coefficients are obtained, then the linking of the separate metrics is performed by applying equations (1) and (2) item by item to the item parameter estimates. To place the person location estimates onto the target metric we apply $\theta_i^* = \zeta(\theta_i) + \kappa$ to each individual's person location or its estimate.

In contrast to linear equating, a second approach, *total characteristic function equating*, uses all the item parameter estimates to determine the values of ζ and κ . The objective in this method (also known as *true score equating*, *test characteristic curve equating*) is to align as closely as possible the initial metric's total characteristic function with that of the target metric. The metric transformation coefficients are the values of ζ and κ that satisfy this objective. This approach requires that the two metrics to be link have either all or a subset of items in common.

As mentioned above, another use of a metric transformation is to transform a metric to make it more meaningful or interpretable. One target metric that has intrinsic meaning for people is the total score metric. For instance, rather than informing a respondent that his or her $\hat{\theta}$ is 1.2, which may or may not have any inherent meaning to the respondent, we can transform the respondent's $\hat{\theta}$ to the more familiar total score metric. That is, a respondent with a $\hat{\theta}$ of 1.2 is told his or her score on the 20-item instrument is 15 (or $15/20 = 0.75$). This transformation is performed through the total characteristic function. The gist of this approach is to sum, across an instrument's items, the response probabilities for a given person location estimate. In a proficiency assessment situation the total score metric indicates the expected number of correctly answered items. A variant of this approach divides this sum by the number of items on the instrument to obtain an expected proportion equivalence for θ (i.e., the proportion of responses of 1), which, in proficiency assessment, is the expected proportion of correct responses. Manuscripts would be most helpful if they included a transformation of scores to an interpretable metric if reporting at an individual level.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 713–723.
- Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345–370.
- De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford.
- De Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory Item Response Models: A generalized linear and nonlinear approach* (pp. 43–74). New York: Springer-Verlag.
- Fraser, C., & McDonald, R. P. (2012). NOHARM 4: A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory [computer program]. Retrieved from <http://noharm.software.informer.com/download>.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Haughton, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16, 342–355.
- Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, 38, 79–93.
- Klauer, K. C., & Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology*, 43, 193–206.
- Linacre, J. M. (2001). *A user's guide to WINSTEPS/MINISTEPS*. Chicago, IL: Winsteps.com.
- Linacre, J. M., & Wright, B. D. (2001). *A user's guide to BIGSTEPS*. Chicago, IL: MESA Press.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing.
- Muraki, E., & Bock, R. D. (2003). PARSCALE (Version 4.1) [computer program]. Mooresville, IN: Scientific Software.
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495–502.
- Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14, 197–207. (A correction may be found in *Applied Psychological Measurement*, 15, 352.)
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271–282.
- Rost, J. (1991). A logistic mixture distribution model for polytomous item responses. *British Journal of Mathematical and Statistical Psychology*, 44, 75–92.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.
- Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333–343.
- Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement*, 45, 433–444.
- Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73–92). Maple Grove, MN: JAM Press.
- Thissen, D. J., Chen, W.-H., & Bock, R. D. (2003). MULTILOG (version 7.0) [computer program]. Mooresville, IN: Scientific Software, Inc.
- Thissen, D. J., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum.
- Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 83–108). New York: Academic Press.
- Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.
- Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *One-parameter logistic model* (OPLM). Arnhem, The Netherlands: CITO.
- Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG (version 3.0) [computer program]. Mooresville, IN: Scientific Software.

12

Latent Class Analysis

Karen M. Samuelsen and C. Mitchell Dayton

Latent class analysis (LCA) is a statistical method for identifying unobserved groups based on patterns of categorical data. LCA is related to cluster analysis (see Chapter 4, this volume) in that both methods are concerned with the classification of cases (e.g., people or objects) into groups that are not known or specified *a priori*. In LCA, cases with similar response patterns on a series of manifest variables are classified into the same latent class with membership in these classes being probabilistic rather than deterministic. LCA can also be viewed as analogous to factor analysis (see Chapter 8, this volume), with the former examining categorical variables and the latter continuous ones; however, this comparison is less direct than in the case of cluster analysis. However, both LCA and factor analysis are based on the principle that observed variables are (conditionally) independent assuming knowledge of the latent structure. Finally, LCA is related to item response theory (see Chapter 11, this volume) and can be viewed as a generalization of discrete response models such as the Rasch model (Lindsay, Clogg, & Grego, 1991).

LCA has been applied in a variety of fields and contexts. In the health and medical fields it has been utilized for: identifying subgroups of risk factors for falling accidents in an elderly population (Hardigan, 2009), distinguishing classes based on changes in intraocular pressure in the context of assessing risk for glaucoma (Gao et al., 2012), the assessment of diagnostic agreement (Uebersax & Grove, 1990), and studying dietary patterns (Patterson, Dayton, & Graubard, 2002). LCA has also been applied to market segmentation research (Wedel & Kamakura, 2000), pedestrian crash injury studies (Sasidharan, Wu, & Menendez, 2015), adolescent gambling research (Goldstein et al., 2013) and cross-cultural studies (Finch & Marchant, 2013).

Overviews of LCA can be found in McCutcheon (1987), Dayton (1999), and Heinen (1996). Related and more complex models are discussed in Hagenaars (1993), Langeheine and Rost (1988), Hagenaars and McCutcheon (2002) and Collins and Lanza (2009). An excellent web-based resource is provided by Uebersax (2000). In addition, selected seminal papers on the subject include: Clogg and Goodman (1984), Dayton and Macready (1976), and Goodman (1974). For an extended bibliography, see the Uebersax website. Finally, as discussed in Desideratum 8, various software programs are available for estimation and analysis of latent class models.

Table 12.1 Desiderata for Latent Class Analysis.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. Substantive theories guiding the choice of models to be evaluated are synthesized.	I
2. Theoretical connections among the manifest variables, covariates, and potential latent classes are explicated.	I
3. The assumption of local independence is discussed and evidence is offered as to why the manifest variables would be independent of each other within latent classes.	I
4. Manifest variables are defined and their appropriateness is justified.	M
5. Categorical and continuous covariates used in the analysis are discussed and a rationale for their inclusion is provided.	M
6. Sampling method(s) and sample size(s) are explicated and justified. The impact of sample size on cell frequencies and model fit statistics should also be discussed.	M
7. The mathematical model(s) being considered are presented along with a substantive justification of the constraints (if any) placed on the allowable values of the parameters to be estimated.	M, R
8. The name and version of the utilized software package are reported. The parameter estimation method is justified and its underlying assumptions are addressed.	M, R
9. Implications of model identification issues for congruence of model with research question(s) are considered.	R
10. Recommended fit indices are presented and evaluated using literature-based criteria.	R
11. For competing models comparisons are made using statistical tests and/or information criteria.	R
12. Evidence-based measures of classification quality are presented and discussed.	R
13. Latent class proportions and conditional probabilities are reported.	R
14. Evidence is provided that global versus local maxima have been reached.	R
15. Boundary value parameter estimates are highlighted and implications are discussed.	R, D
16. Meaningfulness of the latent class proportions is considered.	R, D
17. Membership of the latent classes is discussed.	R, D

* I = Introduction, M = Methods, R = Results, D = Discussion.

1. Theories and Models

Like factor analysis and its generalized formulation, structural equation modeling, LCA uses manifest variables to provide information regarding hypothesized latent variables. In the case of LCA, one or more unobserved, or latent, categorical variables are assumed to be responsible for observed, or manifest, response patterns. LCA can be either exploratory or confirmatory. In the exploratory mode, LCA can be considered to be a special case of cluster analysis with categorical variables. In the confirmatory mode, LCA can be used to fit scaling models to categorical data and, in fact, to fit the equivalent of structural equation models (Hagenaars, 1993). In the absence of an explicit theory regarding the latent variables the researcher is still expected to discuss prior experiences that lead to the choice of latent class models to compare. When in a confirmatory mode, it is incumbent on the researcher to explain, in some depth, the choices made vis-à-vis the models chosen.

In either an exploratory or confirmatory mode, one issue that needs to be addressed is how the hypothesized latent classes represent qualitative differences among subjects. For example, it is possible that there are different latent classes of shoppers and that membership in those classes is predicated on the basis of motivational considerations behind those shoppers'

purchases (e.g., fashion, price, and quality). What experience, literature, or theory supports the existence of these classes?

In the previous example, the latent classes simply represented different categories of shoppers. It is possible that in a different situation the latent classes could be ordered in some way. An example would be latent classes predicated on cigarette smoking behavior, where the latent classes are comprised of initiators, experimenters, regular users, and daily/dependent smokers (Flaherty, 2008). Once again, the question of what experience, literature, or theory substantiates that ordering must be addressed.

Finally, it is possible that one or more unscalable classes might be necessary to adequately model the data (Dayton & Macready, 1980; Goodman, 1975). For example, when young children are asked to select personality statements that best describe themselves, they may have neither a command of the vocabulary needed nor the cognitive understanding when the questions are not concrete in nature (Meijer, Egberink, Emmons, & Sijtsma, 2008). These children will tend to provide inconsistent response patterns because they resort to guessing and will constitute an unscalable class. If one or more unscalable classes are included in a model, the researcher must explain why certain respondents' responses could be inconsistent.

2. Linking Variables and Covariates to Potential Latent Classes

Many important questions in the social sciences involve comparisons between manifest groups. As an example, consider the data concerning academic cheating reported by college students that were examined by Dayton and Scheers (1997). When considering academic cheating it may be interesting to compare groups such as: (1) males and females; (2) juniors and seniors; or (3) students in different academic programs. In the Introduction section of the manuscript, the researcher should discuss why the grouping variables or covariates included in the analysis should be linked to membership in the latent classes. What literature supports this linkage and, if the literature is sufficiently detailed, what differences have been shown to exist between the manifest groups?

3. Local Independence

In LCA, it is assumed that: (1) the model correctly specifies the number of classes, (2) each respondent belongs to only one latent class, and (3) respondents within a class are homogeneous. Building on these, the fundamental concept of LCA is that of local (i.e., conditional) independence meaning that the observed manifest responses are independent given that latent class membership is known. Although, in theory, latent class models may include dependencies among residual terms (e.g., Hagenaars, 1988), this is quite uncommon in the research literature and it should only be done for substantive reasons, not to improve model fit (Flaherty, 2008). In either case, when local independence is assumed or when dependencies among residuals are allowed, the researcher should discuss these assumptions.

4. Manifest Variables

Relatively few assumptions underlie latent class analyses, particularly with regard to the manifest variables under consideration; however, these variables merit some discussion in the Methods section. The most obvious characteristic that should be highlighted is whether they are dichotomous, polytomous, or ordered polytomous in nature. If polytomous variables are employed there should be some justification for the number of scale points used. For example, if manifest variables were based on 5-point Likert scales, the researcher should verify and report that all scale points were

chosen by respondents. If not, it might be appropriate to reduce the number of categories for some or all of the scales or to consider whether or not the variables are more appropriately modeled as dichotomous. Simple descriptive statistics can verify that scale points were used, but need not be shown in the document. If scale points are ordered (as for a Likert scale), then appropriate ordinal modeling methods should be employed (see Rost, 1988).

For confirmatory scaling models (and related models), permissible response patterns for a specific latent class model are also of interest. For example, the Lazarsfeld–Stouffer questionnaire data set (Lazarsfeld, 1950) was based on responses by non-commissioned officers to four dichotomous items regarding attitudes toward the Army. These items were intended to express increasingly more favorable attitudes such that the permissible response vectors were {0000}, {1000}, {1100}, {1110}, and {1111}, where 0 is a negative response and 1 is a positive response. The implication is that when a respondent agrees with the fourth item, one highly favorable toward the Army, the respondent also agrees with the three other items. Theoretically, all other response patterns are not permissible, however this does not mean that they will not occur since response errors and other factors may affect responses (additional discussion of these models is presented in Desideratum 7).

5. Covariates

Concomitant variables can be incorporated into LCA in two different manners: (1) as grouping (stratification) variables; or (2) as covariates that are modeled in a manner similar to logistic regression (see Dayton & Macready, 2002). In either case, some rationale should be provided for the choice of specific concomitant variables and their manner(s) of being included in the latent class model. When grouping variables are used, it is possible to explore homogeneous models (with all parameters constrained equal across groups), partially homogeneous models (with some parameters constrained equal across groups), and heterogeneous models (with no parameters constrained equal across groups). When fully heterogeneous models are not included, some rationale for this omission should be provided. When covariates are used, the form of the regression component of the model (e.g., logistic) should be described. In addition, the rationale for including, or not including, higher-order terms such quadratic functions of continuous covariates should be addressed. Finally, if grouping variables are used as covariates (e.g., by including indicator or dummy variables), the rationale for including, or not including, product terms for continuous and grouping covariates should be addressed. Whenever possible, some form of graphical display for the relation between the covariate(s) and latent class membership should be included.

6. Sampling Method(s) and Sample Size(s)

The inferences that can be drawn from LCA are limited by the sample available for analysis. Samples that are not representative of the population of interest to the researcher will severely limit the inferences that may be drawn from the analysis. Virtually all theoretical treatments of LCA have been based on the notions of simple random sampling, with, perhaps, the inclusion of manifest grouping (stratification) variables. In practice, however, data sets are often based on complex survey designs that involve clusters of respondents and disproportionate sampling. While it is relatively straightforward to incorporate sampling weights to make adjustments for disproportionate sampling, corrections for cluster sampling are more difficult. Failure to take clustering into account can result in severe underestimation of standard errors and can distort results from significance tests. Patterson et al. (2002) discussed these issues and provided some guidance for these situations. In any case, when complex sample designs are involved, these issues should be explicitly discussed in a manuscript.

Sample size is an important consideration in LCA, especially as it relates to observed cell frequencies. For the dichotomous case with V manifest variables, the possible number of unique response vectors is 2^V . While for four variables there are only 16 patterns, for 10 variables there would be 1024 possible response vectors. Thus, with larger numbers of variables, analyses based on frequencies for response vectors are not practical. In addition, goodness-of-fit tests are not applicable unless sample sizes are truly enormous given that frequency tables will be sparse (i.e., contain large numbers of 0 frequencies). In general, analyses with large numbers of manifest variables are possible using raw data (i.e., responses for individual cases rather than frequency data). Note that identification and convergence issues become more critical in these situations and need to be discussed in a manuscript.

7. Mathematical Model

In the notation formalized by Goodman (1974), manifest variables are denoted by capital letters (A , B , C , etc.) and, in the dichotomous case, the variables have levels $i = \{1, 0\}$, $j = \{1, 0\}$, $k = \{1, 0\}$, and so forth, respectively. Latent variables are denoted X and have levels $t = \{1, \dots, T\}$. Therefore, for the t th latent class, the conditional probabilities are represented by $\pi_{it}^{\bar{A}X}$, $\pi_{jt}^{\bar{B}X}$, $\pi_{kt}^{\bar{C}X}$, etc. Assuming three manifest variables and a latent class t , the conditional probability of a response vector y for the s th case associated with the t th class can be written as:

$$P(y_s | t) = \pi_{ijkt}^{\bar{A}\bar{B}\bar{C}X} = \pi_{it}^{\bar{A}X} \cdot \pi_{jt}^{\bar{B}X} \cdot \pi_{kt}^{\bar{C}X} \quad (1)$$

Assuming local independence and latent class proportions π_t^x for a total of T latent classes, the probability of obtaining a response vector can be expressed as the weighted sum across classes:

$$P(y_s) = \sum_{t=1}^T \pi_t^X \cdot \pi_{it}^{\bar{A}X} \cdot \pi_{jt}^{\bar{B}X} \cdot \pi_{kt}^{\bar{C}X} \quad (2)$$

This simple model is generally presented in any article using LCA and provides a frame of reference from which to discuss any constraints placed on the model. These constraints depend upon the purpose of the analysis (exploratory or confirmatory), whether the models under consideration are restricted, and how response errors are modeled (e.g., Proctor, intrusion-omission), as described below.

Typically, exploratory LCA is conducted with no restrictions on the values of conditional probabilities or latent class proportions. In confirmatory analyses, however, the researcher must specify the hypotheses of interest and, typically, constraints are imposed that reflect these hypotheses. For example, the Proctor (1970) scaling model proposes permissible response vectors that represent “true types” in the population and incorporates equal rates of response error for all manifest variables. In a Proctor model with three dichotomous variables (A , B , C) and a linear (i.e., Guttman) scale, the permissible response vectors are $\{000\}$, $\{100\}$, $\{110\}$, and $\{111\}$, reflecting membership in the four latent classes of the latent variable X (classes 1, 2, 3, and 4, respectively). The other possible manifest vectors, $\{001\}$, $\{010\}$, $\{011\}$, and $\{101\}$, reflect some sort of response error. Assuming equal rates of response error for the three variables, as in the Proctor model, the following constraints would be imposed:

$$\begin{aligned} \pi_{11}^{\bar{A}X} &= \pi_{11}^{\bar{B}X} = \pi_{11}^{\bar{C}X} = \\ \pi_{12}^{\bar{B}X} &= \pi_{12}^{\bar{C}X} = \pi_{13}^{\bar{C}X} = \\ \pi_{02}^{\bar{A}X} &= \pi_{03}^{\bar{A}X} = \pi_{03}^{\bar{B}X} = \\ \pi_{04}^{\bar{A}X} &= \pi_{04}^{\bar{B}X} = \pi_{04}^{\bar{C}X} = \pi_{error} \end{aligned} \quad (3)$$

The purpose of this brief explanation of the Proctor model is to show the necessity for discussing the general model in an article. In practice, more complex models may also be investigated that are generalizations of the Proctor model. For example, the intrusion-omission error model (Dayton & Macready, 1976) relaxes the assumption of equal error rates and allows for errors of intrusion (i.e., the occurrence of an observed response of 1 when a permissible vector calls for a 0 response) and errors of omission (i.e., the occurrence of an observed response of 0 when a permissible vector calls for a 1 response).

8. Software Package

There are many good software programs for LCA. Popular ones include Latent Gold (Vermunt & Magidson, 2015), Mplus (Muthén & Muthén, 2012), and WINMIRA2001 (von Davier, 2001). SAS also has two procedures, PROC LCA and PROC LTA, to handle latent class analyses. Note that this is not meant to be an exhaustive list, but rather a sample of the widely used LCA programs. The name and version of the software package must be reported in any manuscript using LCA. This indirectly provides the reader with some information such as whether standard errors are available for parameter estimates, multiple start values are automatically tested, and if local dependence can be handled. All other things being equal, we would recommend using programs that provide standard errors and multiple start values; however the choice of exactly which program to use depends upon many other factors as well (e.g., expense, type of model(s), user experience).

In addition, the parameter estimation method used within the chosen software program, along with its underlying assumptions, should be discussed. Parameters will generally be found using either maximum likelihood (ML) estimation or Bayesian methods such as those available in Latent Gold (Vermunt & Magidson, 2000). Markov chain Monte Carlo (MCMC) methods can also be applied to latent class models; however, this methodology requires special software, original programming, and extensive knowledge of the techniques. Therefore, we would not recommend these methods for anyone but the most advanced users.

9. Model Identification

Most LCA programs use ML estimation which is an iterative procedure that runs until the change in the log likelihood between successive iterations is less than some pre-set criterion. The value of that convergence criterion affects the likelihood that local, rather than global, maxima are identified, and thus this criterion should be reported (see Desideratum 14 for more on local and global maxima).

An identified model is one for which there is a single “best” solution. Non-identified models have multiple “best” solutions. From a practical standpoint a researcher needs to know that the number of parameters to be estimated is limited by the number of unique response vectors minus one, which equals the degrees of freedom. Unfortunately, this is a necessary but not sufficient check for an identified model, because models may be non-identified due to the observed data and their match with a particular model. For this reason it is also common to report that the estimated variance-covariance matrix is of full rank. Note that this matrix contains estimated values for the sampling variances and covariances for the ML parameter estimates.

10. Fit Indices

The goodness of fit of a particular latent class model to the observed data can be assessed using some form of a chi-square test. Two familiar versions of this significance test are the Pearson statistic, χ^2 , and the likelihood ratio statistic, G^2 . The former is based on the differences between observed and

expected frequencies, while the latter is based on the logarithm of the ratio of the observed and expected frequencies. In theory, with dichotomous response variables for an unrestricted full-rank model, the degrees of freedom are:

$$df = 2^V - m - 1 \quad (4)$$

where V is the number of dichotomous manifest variables and m is the number of independent parameters estimated in the model. This equation must be modified appropriately if the rank of the estimated covariance matrix is less than m . This occurs for models with restrictions on the conditional probabilities (e.g., scaling models), for models with conditional probabilities that approach 0 or 1, and for models that are not identified (see Dayton, 1999, for more information regarding these situations). Under most situations, χ^2 and G^2 goodness-of-fit statistics yield similar numerical results and lead to the same conclusions. In general, both statistics are reported along with degrees of freedom. When some cell frequencies are small, both χ^2 and G^2 tend to not follow the appropriate reference chi-square distribution and, thus, become inaccurate. Under this scenario the Read-Cressie statistic is recommended (see Read & Cressie 1988; Dayton, 1999).

One problem with all goodness-of-fit chi-square tests is that, when fitting models based on large sample sizes, the null hypothesis of perfect fit tends to be rejected. This may happen even though the residuals (i.e., differences between observed and expected frequencies) indicate that fit seems satisfactory from a substantive point of view. To complement the limited amount of information provided by chi-square tests, other indices of model fit may be used. Two such indices are the index of dissimilarity, I_D (Dayton, 1999), and π^* (Rudas, Clogg, & Lindsay, 1994). For I_D , defined in terms of observed and expected frequencies, values that are less than 0.05 are generally considered satisfactory. For π^* , which is based on the number of cases that would need to be deleted to achieve perfect model fit, values must be interpreted with reference to the application however 10% is a commonly recommended value (Dayton, 1999). Typical LCA programs do not provide estimates for π^* ; however, a spreadsheet procedure for estimating π^* for two-class models with dichotomous response variables is described and illustrated in Dayton (1999). Given that programs do not provide π^* we recommend that authors include I_D with their results, supplementing with π^* should they choose to.

An indication of how well the model fits the data can also be gained by examining the standard errors, and confidence intervals constructed from them, for the latent class proportions and conditional probabilities. Researchers should, when possible, utilize LCA computer programs that provide estimated standard errors for latent class proportions and conditional probabilities. The accuracy of these estimates has not been widely investigated, especially for small sample sizes; this is one reason for favoring large samples in LCA. If a program is used that does not provide standard errors, it is possible to use resampling methods, such as the jackknife or parametric bootstrap, to estimate standard errors from frequency data.

11. Statistical Tests and Information Criteria

In addition to testing the absolute fit of a model, as discussed in Desideratum 10, it is often of interest to researchers to compare alternate models. For nested models the chi-square difference test can be used in certain circumstances. In those situations one computes the difference between the G^2 statistics for the two models and uses the difference between the degrees of freedom to evaluate whether the more complex model provides statistically significantly better fit. There are, however, technical issues surrounding the use of chi-square difference tests (see Dayton, 1999, for an overview), and for that reason it is often recommended that they are used in concert with measures of relative fit based on information criteria which do not require the models to be nested. Also, it

should be noted that chi-square difference tests cannot be used to compare models based on differing numbers of latent classes (see McLachlan & Peel, 2000 for more on this issue).

Throughout the 1970s and 1980s several information criteria were developed, the first of which was the Akaike information criterion (AIC: Akaike 1973, 1987). This index includes terms for the log likelihood and the number of independent parameters to be estimated, but does not include a term for the sample size and therefore lacks asymptotic consistency (Bozdogan, 1987). Other indices that included sample size parameters are the Bayesian information criterion (BIC: Schwarz, 1978), sample-size adjusted BIC (aBIC: Sclove, 1987) and the consistent Akaike information criterion (CAIC: Bozdogan, 1987). For all of these information criteria, the best fitting model will be the one with the lowest value of the criterion. Often, the strategy is to decide on the preferred model based on multiple information criteria and to report those criteria for all models within a journal article. These criteria are reported in most, if not all, commonly used software programs.

More recently, comparative fit indices have been developed that use approximations to the likelihood ratio distribution to allow for comparisons of models with different numbers of classes. Examples of these tests are the (Vuong)–Lo–Mendell–Rubin likelihood ratio test (VLMR: Lo, Mendell, & Rubin, 2001), the adjusted Lo–Mendell–Rubin test (aLMR: Lo et al., 2001), and the bootstrapped likelihood ratio test (BLRT: MacLachlan & Peel, 2000). These tests can be run in Mplus using the Tech11 and Tech14 commands.

12. Classification Quality

Classification quality can be thought of as the clearness of the delineation between classes. This can be measured three ways; (1) by making comparisons of latent class proportions, (2) using entropy (or relative entropy), and (3) examining the average latent class probabilities for most likely latent class membership. Latent class proportions can be determined based on different methods, some of which may be probabilistic in nature and others based on most likely membership. When these different methods provide the same result there is evidence that the classes are distinguishable. Samuelsen and Raczyński (2013) present an example of comparing information based on the different methods of estimating latent class proportions using the Mplus software.

Information regarding the quality of classifications can also be gleaned by measuring entropy. This statistic, based on the posterior probabilities of class membership, has an upper bound of infinity. For ease of interpretation, relative entropy is often reported (Mplus uses this statistic) which has been rescaled to be bounded by 0 and 1. When entropy values approach 1.0 there is evidence that a clear delineation between the latent classes exists (Celeux & Soromenho, 1996). Entropy, as a single number, provides no information on which classes are the least delineated. In contrast, examining the comparisons of the latent class proportions will provide more detailed information. Specifically, the magnitude of the off-diagonal cells in a table of the average latent class probabilities by latent class can provide evidence of which classes are the most and least distinguishable.

It should be noted that measures of classification quality are not necessarily measures of model fit. Rather, they provide the researcher with an indication of how useful a model might be for a given purpose. If a researcher has no substantive reason that latent classes should be distinct from one another, then relatively low entropy scores or higher average latent class probabilities on the off diagonal would not be indicators that the model did not fit the data.

Several authors (Muthén & Asparouhov, 2006; Skrondal & Rabe-Hesketh, 2004) have suggested that the standardized Pearson residuals for commonly occurring response patterns can provide evidence of classification quality. Specifically, standardized residuals for individual and pairs of items can be compared to a standard normal distribution, with those greater than 1.96 being problematic. When making model comparisons, researchers can compare the number of statistically significant

Table 12.2 Latent Class Structure for Two Hypothetical Items.

<i>Response Pattern</i>	<i>Frequency</i>	<i>Conditional Probabilities</i>	
		<i>Latent Class 1</i>	<i>Latent Class 2</i>
{0 0}	516	0.004	0.512
{1 0}	144	0.016	0.128
{0 1}	164	0.036	0.128
{1 1}	176	0.144	0.032
Total	1000	0.200	0.800

univariate and bivariate residuals across models. Though this would provide evidence regarding which model has higher classification quality than the other, there are no rules of thumb that allow researchers to say that an individual model is acceptable.

13. Latent Class Proportions and Conditional Probabilities

When relatively few response patterns are possible, it is helpful to present those patterns, the conditional probabilities, and the latent class proportions in one table. The frequency of each response pattern and total sample size can also be reported. As shown in Table 12.2 (adapted from Dayton, 1999), arranging data in this manner provides the reader with a comprehensive summary of the analysis in question.

When the number of unique response patterns becomes large it is more convenient to show how the responses to individual items differ across the classes. In the example in Table 12.3 (adapted from Dayton, 1999), four dichotomous items (*A*, *B*, *C*, and *D*) were examined. The conditional probabilities for both the positive response (1) and negative response (0) are shown even though that information is redundant. In this case it is obvious that the first latent class (labeled LC 1) was much larger than the second (approximately 85% to 15%), that for LC 1 there are boundary value estimates (see Desideratum 15) for variable *B*, and that the conditional probabilities of a positive response differ substantially across classes for all of the items. The standard errors included in parentheses also indicate to the reader which parameters were estimated with greater precision. A similar table could be created for polytomous items with zero through the number of response options shown for each item. This representation could also be used to show similarities and differences among manifest groups.

Table 12.3 Cheating Data for Males Only (Standard Errors in Parentheses).

		<i>Latent Class 1</i>	<i>Latent Class 2</i>
		0.8545 (0.0595)	0.1455 (0.0595)
<i>A</i>	0	0.9775 (0.0281)	0.4299 (0.1702)
<i>A</i>	1	0.0225 (0.0281)	0.5701 (0.1702)
<i>B</i>	0	1.0000 (0.0000)	0.5486 (0.1943)
<i>B</i>	1	0.0000 (0.0000)	0.4514 (0.1943)
<i>C</i>	0	0.9944 (0.0184)	0.6817 (0.1341)
<i>C</i>	1	0.0056 (0.0184)	0.3183 (0.1341)
<i>D</i>	0	0.8553 (0.0364)	0.5457 (0.1425)
<i>D</i>	1	0.1447 (0.0364)	0.4543 (0.1425)

For exploratory analyses when models are estimated based on differing numbers of classes, a simple graphical display of the conditional probabilities can be included. A rationale for why these graphs are valuable can be seen in Figure 12.1 from a six-item dataset regarding attitudes toward abortion (Dayton, 2006). The first three items (labeled 1, 2, and 3 on the x -axis) deal with favoring abortions for reasons such as mother's health, rape or incest, and birth defect, whereas the last three items (labeled 4, 5, and 6) deal with favoring abortions for reasons such as being unmarried, being poor or having too many children already. The first class in the 4-class model includes individuals who are strongly pro-abortion without restriction. The individuals in the second and third classes put conditions on their support of abortion with relatively strong support for items 1–3 but weaker support for items 4–6. The fourth class includes individuals who express fairly negative opinions about abortion regardless of the situation. The 5-class model shows these same four classes plus another class that seems to include individuals who may be responding randomly to the items. Given that the responses do appear random, and that this class only represents 1% of the respondents, it would be appropriate to choose the 4-class model over the 5-class model even if the more complex showed better fit.

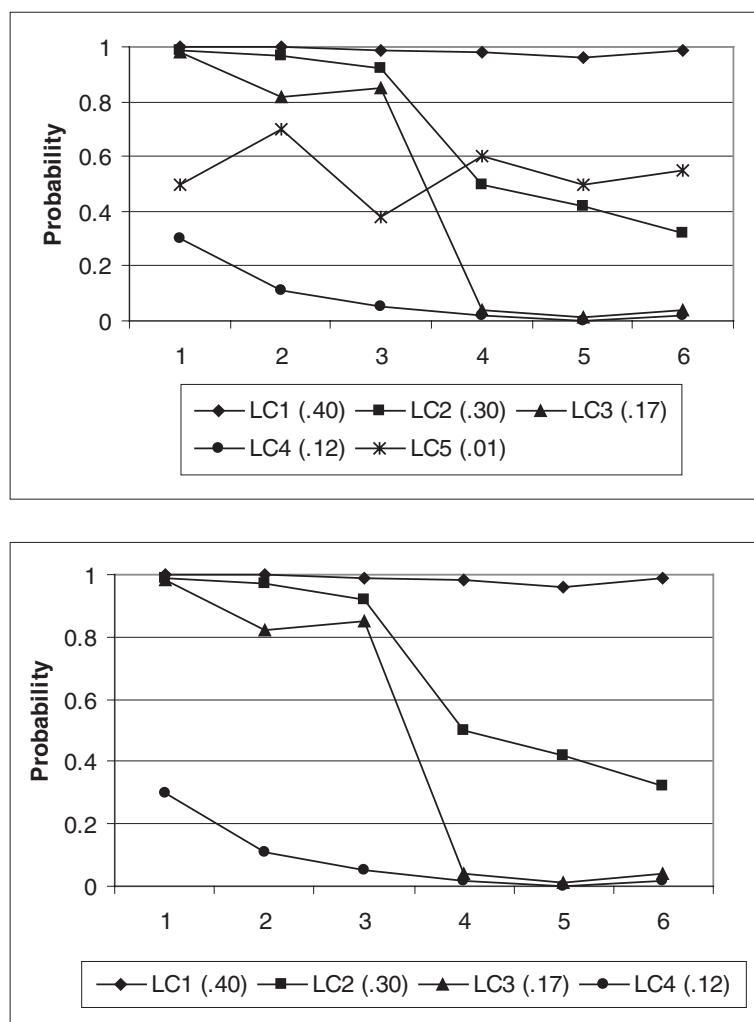


Figure 12.1 Profiles for 4- and 5-Class Models.

14. Global vs. Local Maxima

Because of the mathematical form of the equations that must be solved to conduct maximum-likelihood estimation with latent class models, the ordering of latent classes is arbitrary. For example, for a two-class solution where the first latent class proportion is .68 there is an equivalent two-class solution where the first class proportion is .32 (and the item conditional probabilities are correspondingly switched). In the technical literature, this is known as *label switching*. Although, in practice, this is merely a bookkeeping issue, it must be kept in mind when comparing solutions with varying number of latent classes or solutions with different sets of restrictions given that the classes may not directly correspond across different solutions.

Another issue is that local maxima might exist. Ideally, the algorithm being used to estimate model parameters should seek out the global maximum. However, existing algorithms using ML or Bayesian estimation cannot, in general, distinguish between the global maximum and locally optimal solutions. Thus, when reporting LCA solutions it is necessary to provide evidence that, in fact, a global maximum was reached. There are several types of evidence that are suitable. First, the LCA program should be executed several times with different sets of start values. If all of these runs result in the same solution, this is strong evidence that the global maximum has been reached. If those runs reach different maxima, it may be defensible to select an analysis with the largest log-likelihood but this requires explicit discussion as it is possible that the model is not identified. An article should also include the minimum convergence criteria used; stringent criteria (e.g., the change in the log-likelihood between iterations is 10^{-8} or less) will lend more credence that a global maximum has been reached when multiple LCA solutions identify that same maximum. Parameters whose estimates approach boundary values of 0 or 1 should be cause for suspicion. The greater the number of parameter estimates at the boundary values, the greater a researcher should be concerned that the solution is a local, rather than global, maximum. Desideratum 15 provides a more in-depth discussion of boundary values.

15. Boundary Value Parameter Estimates

When estimated conditional probabilities approach 0 or 1, computational issues arise with respect to estimated variances and covariances for the parameter estimates that render the associated confidence intervals and significance tests of questionable meaning (Garre & Vermunt, 2006). Therefore, in the Results section, one should be very specific regarding which conditional probabilities, if any, went to boundary values of 0 or 1.

Because estimates approaching boundary values might indicate that the data have been overfit, the researcher generally seeks to simplify or restrict the model in some way. One recommendation is to restrict these parameters to 0 or 1 (as appropriate), re-estimate the remaining parameters, and reduce the degrees of freedom for chi-square goodness-of-fit tests by 1 for each such restriction.

Certain LCA programs, notably Latent Gold (Vermunt & Magidson, 2015), incorporate Bayesian methods with Dirichlet prior probabilities for latent class parameters so that the boundary value issue is, for practical purposes, eliminated. Also, there is evidence that this assists in avoiding local maxima without distorting results. In general, Bayesian estimates will not be the same as ML estimates for identified models. For example, for identified models with the weak priors incorporated as defaults in Latent Gold, Bayesian parameter estimates tend to be “shrunken” relative to conventional ML estimates (i.e., tend to be further from the 0/1 boundaries). Therefore it is important to specify the rationale for using Bayesian methods and to describe the prior distributions being utilized.

16. Meaningfulness of Latent Class Proportions

Researchers need to not only consider which models fit best based on statistical indices, but on the sizes of the classes as well. This is especially true if one of the identified classes is extremely small; however, the definition of “small” depends upon context. For example, with a sample size of 200, a latent class proportion of .05 represents only 10 cases, whereas with a sample size of 20,000 it represents 1000 cases. In essence, the question is whether or not a class is substantively meaningful or an artifact of the particular sample being analyzed. When a latent class proportion is extremely small, it is incumbent upon the researcher to offer a substantive rationale, based on past research or theory, for the inclusion of the class.

17. Latent Class Membership

A posteriori probabilities for each response vector and latent class can be computed and used to classify respondents into one of the latent classes. The specificity of the discussion regarding latent class membership will depend on the number of response vectors and how those correspond with the latent classes. If a few different response vectors were responsible for latent class membership it would be informative to state exactly what those vectors were. For example, with academic cheating data for four items, Dayton and Scheers (1997) found that all response vectors, except {0000}, {0010}, {0001}, and {0011}, resulted in classification in a latent class identified as representing persistent cheaters. It appears, then, that classification as a non-persistent-cheater hinged on responding “no” (i.e., 0) to the first two items. In situations with more items, on the other hand, if dozens of response vectors were associated with each class it would be of little value to enumerate those vectors. Instead, it would be more informative to discuss the items that seemed to be most predictive of class membership.

The interpretation of latent classes can be enhanced by post hoc analyses that involve concomitant variables. This may be viewed as a type of construct validation for the naming of latent classes. For example, the analysis of academic cheating data by Dayton and Scheers (1997) incorporated student grade point average as a covariate and resulted in an interpretation of the two classes as persistent cheaters and non-persistent-cheaters. Note that there is a substantive rationale for assuming the probability of respondents being classified in these latent classes would be related to academic success in college.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csake (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.
- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52, 317–332.
- Albert, P. S., McShane, L. M., & Shih, J. H. (2001). Latent class modeling approaches for assessing diagnostic error without a gold standard: With applications to p53 immunohistochemical assays in bladder tumors. *Biometrics*, 57, 610–619.
- Bozdogan, H. (1987). Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345–370.
- Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13, 195–212.
- Clogg, C. C., & Goodman, L. A. (1984). Latent structure analysis of a set of multi-dimensional contingency tables. *Journal of the American Statistical Association*, 79, 762–771.
- Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. Hoboken, NJ: Wiley.
- Dayton, C. M. (1999). *Latent class scaling analysis*. Thousand Oaks, CA: Sage.
- Dayton, C. M. (2006). Latent structure of attitudes toward abortion. In S. S. Sawilowsky (Ed.), *Real data analysis* (pp. 293–298). Greenwich, CT: Information Age Publishing.
- Dayton, C. M., & Macready, G. B. (1976). A probabilistic model for validation of behavioral hierarchies. *Psychometrika*, 41, 189–204.

- Dayton, C. M., & Macready, G. B. (1980). A scaling model with response errors and intrinsically unscalable respondents. *Psychometrika*, 45, 343–356.
- Dayton, C. M., & Macready, G. B. (2002). Use of categorical and continuous covariates in latent class analysis. In A. McCutcheon & J. Hagenaars (Eds.), *Advances in latent class modeling* (pp. 213–233). Cambridge: Cambridge University Press.
- Dayton, C. M., & Scheers, N. J. (1997). Latent class analysis of survey data dealing with academic dishonesty. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 172–180). Munich: Waxman Verlag.
- Finch, W. H., & Marchant, G. J. (2013). Application of multi-level latent class analysis to identify achievement and socio-economic typologies in the 20 wealthiest countries. *Journal of Educational & Developmental Psychology*, 13, 201–221.
- Flaherty, B. P. (2008). Examining contingent discrete change over time with associative latent transition analysis. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 299–316). Charlotte, NC: Information Age Publishing.
- Gao, F., Miller, J. P., Miglior, S., Beiser, J. A., Torri, V., Kass, M. A., & Gordon, M. O. (2012). The effect of changes in intraocular pressure on the risk of primary open-angle glaucoma in patients with ocular hypertension: An application of latent class analysis. *BMC Medical Research Methodology*, 12, 151.
- Garre, F. G., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika*, 33, 43–59.
- Gelfand, A. E., & Solomon, H. (1974). Modeling jury verdicts in the American legal system. *Journal of the American Statistical Association*, 69, 32–37.
- Goldstein, A. L., Faulkner, B., Cunningham, R. M., Zimmerman, M. A., Chermack, S., & Walton, M. A. (2013). A latent class analysis of adolescent gambling: Application of resilience theory. *International Journal of Mental Health & Addiction*, 7, 13–30.
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215–231.
- Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association*, 70, 755–768.
- Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators: Local dependence models. *Sociological Methods & Research*, 16, 379–405.
- Hagenaars, J. A. (1993). *Log-linear models with latent variables*. Thousand Oaks, CA: Sage.
- Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press.
- Hardigan, P. C. (2009). An application of latent class analysis in the measurement of falling among a community elderly population. *The Open Geriatric Medicine Journal*, 2, 12–17.
- Heinen, T. (1996). *Latent class and discrete trait models*. Thousand Oaks, CA: Sage.
- Langeheine, R., & Rost, J. (Eds.). (1988). *Latent trait and latent class models*. New York: Plenum Press.
- Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. R. Lazarfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). Princeton, NJ: Princeton University Press.
- Lindsay, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86, 96–107.
- Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88, 767–778.
- McCutcheon, A. C. (1987). *Latent class analysis*. Thousand Oaks, CA: Sage.
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
- Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment*, 90, 227–238.
- Muthén, B., & Asparouhov, T. (2006). Growth mixture analysis: Models with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Advances in longitudinal data analysis* (pp. 143–165). Boca Raton, FL: Chapman & Hall/CRC Press.
- Muthén, L. K., & Muthén, B. O. (2012). Mplus User's Guide (Version 7). Los Angeles, CA: Muthén and Muthén.
- Patterson, B., Dayton, C. M., & Graubard, B. (2002). Latent class analysis of complex survey data: Application to dietary data. *Journal of the American Statistical Association*, 97, 721–729.
- Proctor, C. H. (1970). A probabilistic formulation in statistical analysis of Guttman scaling. *Psychometrika*, 35, 73–78.
- Read, T. R. C., & Cressie, N. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer-Verlag.
- Rost, J. (1988). Rating scale analysis with latent class models. *Psychometrika*, 53, 327–348.
- Rudas, T., Clogg, C. C., & Lindsay, B. G. (1994). A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society, Series B*, 56, 623–639.
- Samuelsen, K. M., & Raczyński, K. (2013). Latent class/profile analysis. In Y. Petscher, C. Schatschneider, & D. Compton (Eds.), *Applied quantitative analysis in education and the social sciences* (pp. 304–328). New York: Routledge.
- Sasidharan, L., Wu, K.-F., & Menendez, M. (2015). Exploring the application of latent class cluster analysis for investigating pedestrian crash injury severities in Switzerland. *Accident Analysis & Prevention*, 85, 219–228.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.
- Uebersax, J. (2000). A brief study of local maximum solutions in latent class analysis. Retrieved February 9, 2016, from <http://john-uebersax.com/stat/local.htm>.
- Uebersax, J. S., & Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine*, 9, 559–572.
- Vermunt, J. K., & Magidson, J. (2015). Upgrade manual for Latent GOLD 5.1. Belmont, MA: Statistical Innovations, Inc. Retrieved February 9, 2012 from: www.statisticalinnovations.com.
- Von Davier, M. (2001). *WINMIRA user manual*. Kiel, Germany: Institut für die Pädagogik der Naturwissenschaften.
- Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Dordrecht, The Netherlands: Kluwer.

13

Latent Growth Curve Models

Kristopher J. Preacher

Structural equation modeling (SEM) is one of the most flexible and commonly used tools in the statistical toolbox of the social scientist. *Latent growth curve modeling* (LGM), the subject of this chapter, is one application of SEM to the analysis of change. In LGM, repeated measures of a variable (hereafter, Y) are treated as indicators of latent variables, called *basis curves*, that represent aspects of change—typically intercept and linear slope factors. Values of the time metric (e.g., age, day, or wave of measurement) are built into the factor loading matrix to reflect the form of the hypothesized *trajectory*, or trend over time. There are many extensions of this idea, but these are the basic elements common to all applications of LGM. LGM contains elements of both variable-centered and person-centered approaches (Sterba & Bauer, 2010a,b), in that a sample-level summary of change is provided, yet individual differences in initial status and change are also considered.

The advantages of LGM over other techniques for modeling change are numerous. A primary advantage is that LGM affords the researcher the ability to model aspects of change as *random effects*; that is, the means, variances, and covariances of individual differences in intercepts and slopes can be estimated. Because LGM is a special case of SEM, all of the benefits of SEM apply to LGM as well. These include the ability to assess the fit of the model to data, the ability to assess change in latent variables, and the ability to examine the antecedents and sequelae of change. Missing data (assuming they are missing at random) pose no problem for LGM. Perhaps the greatest advantage of LGM is its flexibility. Cases need not be measured at the same occasions, or even at equally spaced intervals. Complex nonlinear trajectories can be modeled. LGM can be adapted in creative ways to address new problems.

There are currently three book references devoted exclusively or primarily to LGM (Bollen & Curran, 2006; Duncan, Duncan, & Strycker, 2006; Preacher, Wichman, MacCallum, & Briggs, 2008). In addition, there exist many other accessible sources on the subject (e.g., Byrne & Crombie, 2003; Chan, 1998; Curran, 2000; Curran & Hussong, 2003; Hancock, Harring, & Lawrence, 2013; Little, 2013; McArdle, 2012; McArdle & Nesselroade, 2014; Singer & Willett, 2003; Willett & Sayer, 1996). Any SEM software capable of accommodating mean structures and multiple groups (e.g., AMOS, EQS, lavaan, LISREL, Mplus, OpenMx, SAS PROC TCALIS, Stata) may be used to specify these models.

Because LGM is a special application of SEM, the reader may notice some degree of overlap between the desiderata enumerated here and those described in Chapter 33 of this volume. Table 13.1 of the present chapter addresses some of these in the LGM context, and includes

Table 13.1 Desiderata for Latent Growth Curve Models.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. Substantive theories motivating the model under scrutiny are described; a set of <i>a priori</i> specified competing models is generally preferred.	I
2. The metric of time (or, more generally, the substrate of change) should be reported.	I
3. The functional form of the hypothesized trajectory of change is delineated.	I
4. Path diagrams are presented to facilitate the understanding of the conceptual model of change and the specification of the statistical model.	I
5. The scope of the study is outlined; if the author(s) detailed a theory of longitudinal change, enough time must be permitted to elapse for the phenomenon of interest to unfold.	I
6. Repeatedly measured variables are defined and their appropriateness for inclusion in the study is justified.	M
7. The integration of theoretically relevant control variables into the model is explained.	M
8. The sampling method(s) and sample size(s) are explicated and justified.	M
9. The treatment of missing data and outliers is addressed.	M, R
10. The name and version of the utilized software package are reported; the parameter estimation method is justified and its underlying assumptions are addressed.	M,R
11. Problems with model convergence, offending estimates, and/or model identification are reported and discussed.	R
12. Summary statistics for measured variables are presented; if raw data were analyzed, information on how to gain access to data is provided.	R
13. Several model fit indices of multiple types are presented and evaluated using literature-based criteria.	R
14. If incremental fit indices are used, an appropriate null model is specified and fit rather than relying on incorrect estimates provided by software.	R
15. For competing models, comparisons are made using statistical tests (for nested models) or information criteria (for non-nested models).	R
16. For any post hoc model re-specification, theoretical and statistical justifications are provided and the model is fit to a new sample.	R
17. Parameter estimates, together with information regarding their statistical significance, are provided.	R
18. Appropriate language regarding model tenability and structural relations is used.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

additional desiderata specific to LGM. To get the most out of this chapter, it should be read in conjunction with Chapter 33.

1. Substantive Theories and Latent Growth Curve Models

LGM is intended as a way to test the *a priori* predictions of a theory of change against observed data. Therefore, it is critical that the researcher have a well-articulated theory of change before attempting to use LGM. The Introduction section of an article using LGM should build a case for testing specific hypotheses of change. Typically, this involves stating a theoretical reason for specifying individual trajectories that are characterized by aspects of change (intercept, slope, and so on) that, in turn, are expected to vary across sampling units. The point of most applications of LGM is to obtain estimates of the means, variances, and covariances of these trajectories, and

these parameters should have consequences for the theory under scrutiny. The lack of strong theoretical predictions can lead to the misuse of LGM to generate theory from data in an exploratory, inductive fashion. As with SEM in general, testing alternative models of change provides a more comprehensive survey of competing ideas—and is more scientifically sound—than testing a single model of change.

Given that the researcher has in mind a strong theory of change that makes LGM an appropriate analytic strategy, then attention must be given to the question “change in what?” Most applications of LGM involve modeling change in the same variable over time, but this is a questionable undertaking if the nature or meaning of the variable itself changes over time. Change in the fundamental meaning of the variable could be mistaken for change in mean level. For example, common age-appropriate aggressive behaviors in 6-year-old children may decrease in frequency over time not because levels of the aggression construct decline, but rather because other aggressive behaviors take their place. One way to address this phenomenon is by invoking theory or past findings to support the stability in interpretation of the repeatedly measured variable. Another way is to replace directly observed repeated measures with repeated latent variables, each of which has multiple indicators at each measurement occasion. This approach permits tests of *longitudinal factorial invariance*, a way to assess stability in the meaning of a construct over time. Applications that use repeated latent variables should explicitly address longitudinal invariance.

2. The Metric of Time

The overwhelming majority of applications of LGM involve some metric of time as the substrate of change. Time can be measured in units ranging from milliseconds to decades. Quasi-time metrics such as wave of study, developmental stage, or school grade may be used. In fact, the data analyzed with LGM need not be longitudinal at all. For example, there is theoretically no hindrance to replacing the time metric with, for example, stimulus intensity, distance, or dosage, assuming these repeated measures are assessed or administered in a within-subject fashion and ordered in some logical way. In the case of stimulus intensity, the “origin” measure (typically time 0 in a longitudinal study) could represent the absence of a stimulus; it could represent the baseline dosage of a drug in a repeated-measures medical trial. For the remainder of this chapter I refer to the metric as “time,” but this is not meant to exclude other substrates of change.

Two factors should be considered in any application of LGM: *origin* and *scale*. The origin of the time metric refers to the zero-point. The location of the zero-point (e.g., age 37, initial wave of measurement, or time of intervention) has implications for the interpretation of model parameters related to the intercept. For example, the first occasion of measurement is often chosen as the origin to permit interpretation of the intercept as “initial status,” although other choices are feasible and more appropriate in different circumstances (e.g., “time of death” as the last measurement occasion). The scale of a metric refers to the unit of time (e.g., year, minutes since treatment, or developmental stage). Scale is important mainly for interpreting parameters related to the slope, because (linear) slopes are interpreted as the model-implied change in the outcome per unit increase in time. The choice of origin and scale either should be justified by the researcher or should be obvious from the context, and should not be chosen arbitrarily.

3. Functional Form

Most applications of LGM involve testing linear trends. That is, the researcher hypothesizes that scores on the repeated measures proceed upward or downward in a linear fashion. Individuals may vary around this mean linear trend if intercept and slope variances are also estimated, meaning that

the basic growth curve model provides for variability in level and rate of change. In LGM, values representing the trend are entered into columns of a matrix of factor loadings (Λ) in the following way. The first column of Λ always consists of a column of 1s to act as multipliers for the intercept factor. The remaining columns represent functions of the values of the time metric. For example, for a simple linear trend with four equally spaced repeated measurements, Λ could be represented in any of the following equivalent ways:

$$\Lambda_A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \quad \Lambda_B = \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 2 \\ 1 & 4 \end{bmatrix} \quad \Lambda_C = \begin{bmatrix} 1 & -3 \\ 1 & -2 \\ 1 & -1 \\ 1 & 0 \end{bmatrix} \quad \Lambda_D = \begin{bmatrix} 1 & 10 \\ 1 & 20 \\ 1 & 30 \\ 1 & 40 \end{bmatrix},$$

where the first column is always the constant 1 and the second column (containing actual values of the time metric) represents linear growth. The factor loadings in the Λ matrix often are represented in diagram form as path values on arrows connecting intercept and slope factors to the repeated measures of the outcome variable. In Figure 13.1, the loadings in Λ_A , Λ_B , Λ_C , and Λ_D are depicted in simplified path diagram form (path diagrams are treated at greater length under Desideratum 4). Notice that the location of the zero-point, and thus the occasion at which the intercept is interpreted, changes from specification to specification, as does the metric of time. The choice of both the origin and scale of the time metric should be consistent with theory and with the research context. The origin should be chosen carefully to correspond with a theoretically important occasion (e.g., the time of initial assessment, time of intervention, or time of death) so that time can be thought of as *time since* (or until) that event. In the first loading matrix (Λ_A), the origin is placed at the first occasion of measurement. In the second (Λ_B), the origin occurs at the second occasion of measurement, and so on. The scale is always chosen to correspond to a theoretically important metric. For example, change over time in Λ_B might be measured in two-month intervals, with an intervention

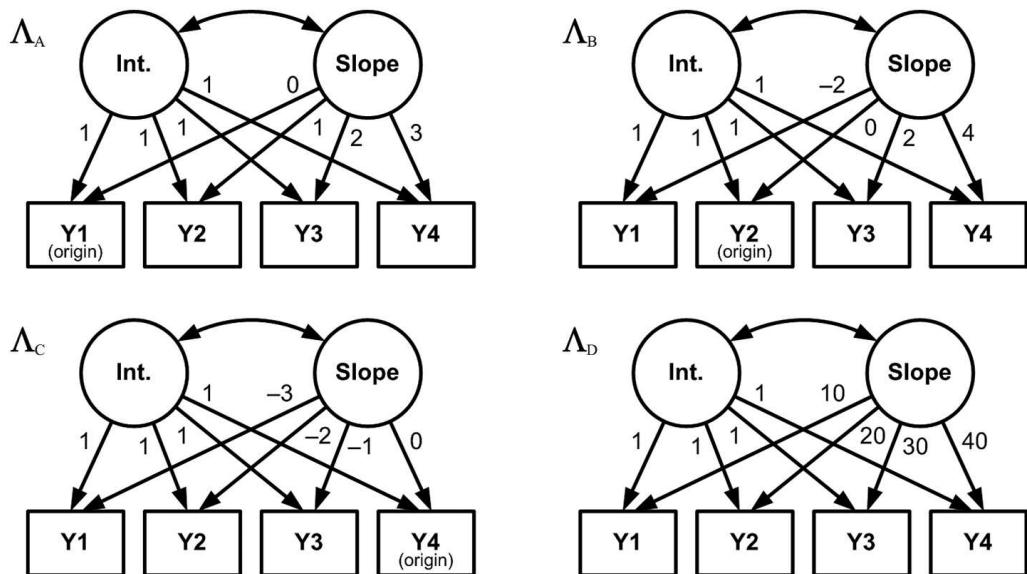


Figure 13.1 How the Loadings in Λ_A , Λ_B , Λ_C , and Λ_D Might Be Represented in Path Diagrams.

imposed at the second occasion of measurement, whereas in Λ_D the unit of time is age in years, and the origin (i.e., birth) falls 10 years before the first assessment. It is usually not necessary to explicitly provide the Λ matrix, especially if an appropriately labeled path diagram is provided (see Desideratum 4), but it often can be helpful as an aid to understanding.

The basic linear LGM can be extended in numerous creative ways. For example, the first Λ matrix below (Λ_E) specifies quadratic growth for four equally spaced repeated measurements—the first column provides loadings for an intercept factor, the second for a linear factor, and the third for a quadratic factor. The Λ_F loading matrix provides for linear growth over four unequally spaced measurement occasions (the second occasion being 4 time units since the initial occasion, the third 5 units, and the fourth 7 units). The Λ_G loading matrix represents an unspecified trajectory, in which the researcher has no specific trajectory in mind, but is willing to let the data determine the shape of change over time.

$$\Lambda_E = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} \quad \Lambda_F = \begin{bmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 5 \\ 1 & 7 \end{bmatrix} \quad \Lambda_G = \begin{bmatrix} 1 & 0 \\ 1 & \lambda_{2,2} \\ 1 & \lambda_{3,2} \\ 1 & 1 \end{bmatrix}$$

There are two major points that should be addressed concerning functional form. The first is that neither the functional form of the hypothesized trajectory nor the measurement schedule needs to conform to a rigid and limited set of options. Flexibility is a hallmark of LGM. The second point is that, regardless of what trend is hypothesized and fit, it must be explicitly justified on the basis of theory. It is rarely a good idea to use LGM in a theoretical vacuum, or to use it to approximate a trend of unknown shape for descriptive purposes. The option to approximate a functional form rather than test an *a priori* hypothesized one does exist, as the Λ_G matrix above demonstrates, but this practice is exploratory, not confirmatory, so conclusions should be worded to reflect the partly atheoretical nature of such trends.

4. Path Diagrams

The use of *path diagrams* is explained in Chapter 33 of this volume. Everything said about path diagrams in Chapter 33 applies here as well, because LGM is a special case of SEM. Path diagrams are not required in applications of LGM, but they almost always greatly facilitate interpretation, especially for readers unacquainted with the method.

Very spartan diagrams were used to illustrate loadings under Desideratum 3. An example of a full latent growth curve path diagram is given in Figure 13.2. As in other SEM path diagrams, circles represent latent variables, squares are measured variables (here, repeated measures), single-headed arrows are path coefficients (regression-type weights), and double-headed arrows are variances or covariances. Aspects of change (intercepts, slopes, and so on) are considered latent variables because they cannot be directly observed. They usually are permitted to vary across people and to covary with one another (e.g., initial status may covary with rate of change), so parameters representing those variances and covariances are often included in the diagram (therein labeled as ψ). Together, the Intercept and Slope factors comprise the *latent trajectory*. In addition to the information provided in Chapter 33 in this volume, there are several noteworthy features specific to diagrams used in LGM. The triangle represents a constant 1.0, and is otherwise treated as a variable. Therefore, the path coefficients labeled as α_1 and α_2 represent the means of the Intercept and Slope latent variables, respectively. Occasion-specific residual variances, labeled as $\theta_{\varepsilon<(1-5)}$ in Figure 13.2, are included to

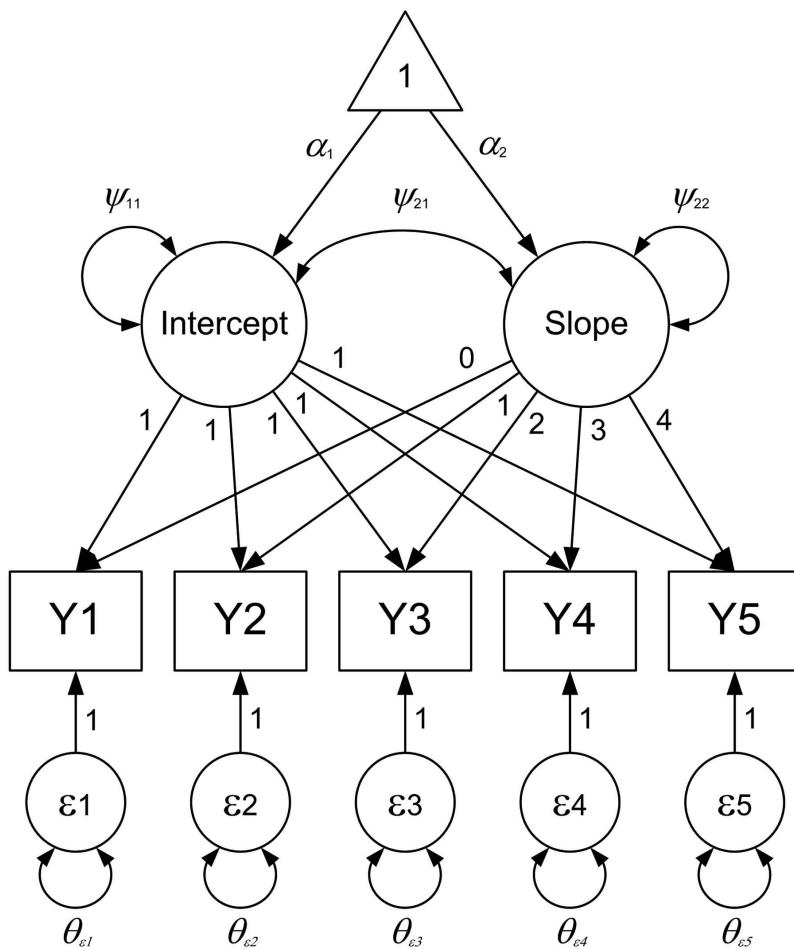


Figure 13.2 Path Diagram of a Linear Latent Growth Curve Model with Random Intercepts, Random Slopes, and Unconstrained Residual Variances.

represent the portion of the variance in the outcome (at a particular occasion) not explained by the latent trajectory.

The other noteworthy feature of Figure 13.2 is the set of loadings connecting the latent trajectory factors with the outcomes. Unlike most applications of SEM or confirmatory factor analysis (see Chapter 8), all of these factors are typically fixed to point values that reflect the hypothesized trajectory. In Figure 13.2, that trajectory is linear, but that can be changed by altering the elements of the Λ matrix (see Desideratum 3). If any elements of Λ are freely estimated, the researcher is approximating an unknown trend rather than testing a hypothesis about a specific trend.

Growth curve models do not have to conform to the linear model depicted in Figure 13.2. For example, more latent trajectory factors may be added (e.g., quadratic, cubic). The equality constraint on residual variances can be freed, within certain limits (see Desiderata 5 and 11). The observed repeated measures can be replaced with latent variables. Because LGM is a special application of SEM, Intercept and Slope factors may serve as independent or dependent variables in larger path diagrams. Multiple latent growth curves may be included in the same model. Figure 13.3 depicts an elaborate example of hypothesized linear growth in a latent aggression construct from ages 6 through 9,

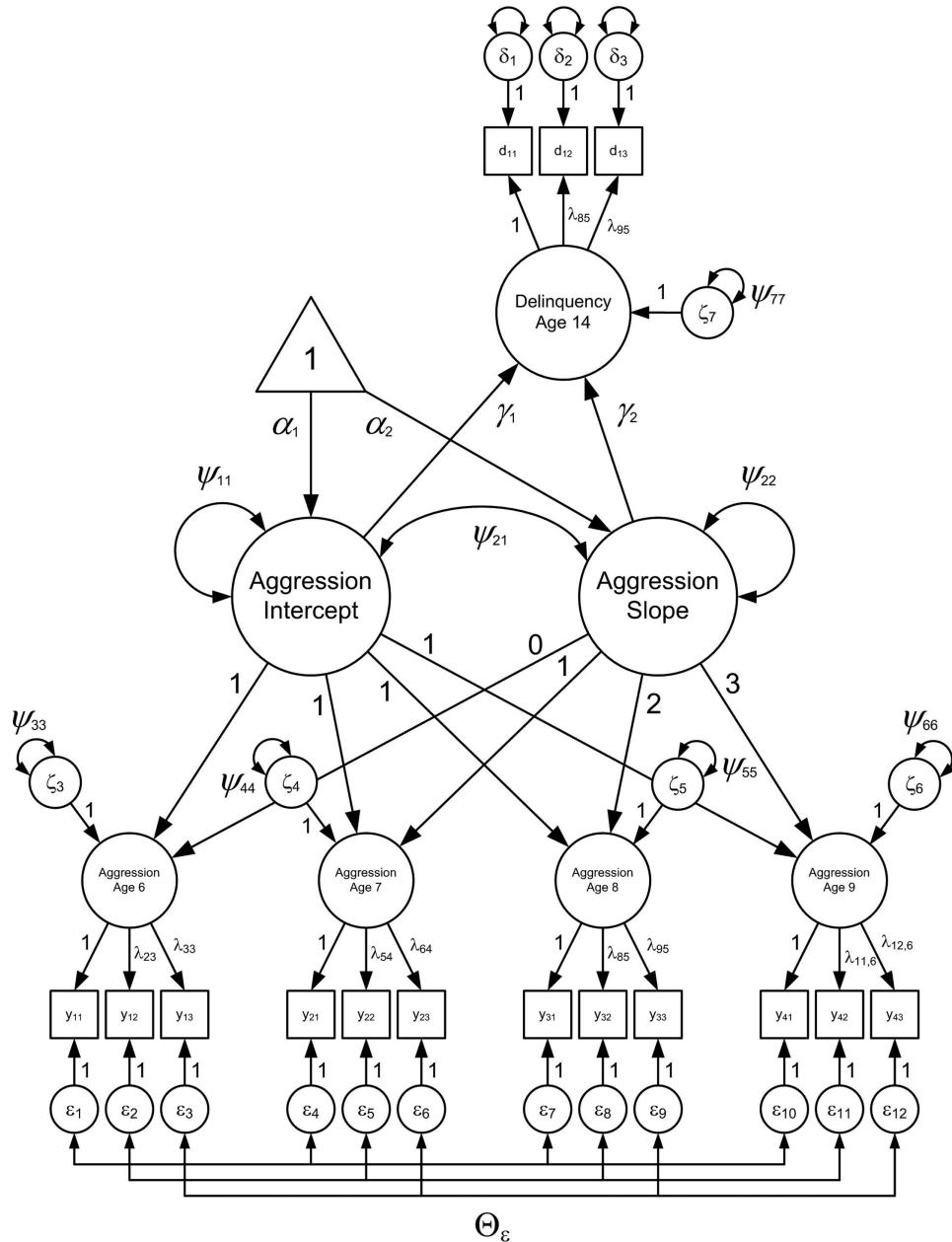


Figure 13.3 A Linear LGM Modeling Growth in Repeated Measures That Are Themselves Latent Variables with Multiple Indicators.

where the effects of Intercept and Slope on Delinquency at age 14 are of interest. Residuals of similar indicators (i.e., those of y_{11}, y_{21}, y_{31} , and y_{41}) are permitted to covary in this example.

5. Scope of the Study

Here, *scope* refers to the number and range of repeated measurements. LGM finds its greatest use when there are only a few repeated measurements per case (but usually at least 4 or 5), and when the

sample size is large (see Desideratum 8), although LGM is by no means limited to such situations. The three most important questions to consider about the scope of a study are:

1. *Were a sufficient number of occasions chosen to identify the model?* A model is *identified* if the type and number of constraints is sufficient to guarantee a unique solution for every model parameter. For time-balanced data (in which data are collected over a limited number of discrete occasions), there must be at least $k + 1$ repeated measures for the model to be identified, where k is the number of basis curves (i.e., growth trajectory factors). This rule applies regardless of whether or not the residual variances are constrained to equality over time. For example, because a cubic trend requires four basis curves (intercept and linear, quadratic, and cubic slope components), there must be at least five repeated measures. If there are exactly five repeated measures and residual variances are constrained to equality, then $df = 5$. If residual variances are freely estimated, then $df = 1$. This rule is simply an elaboration of the assertion that two points define a line. At least two repeated measures are required in order for linear growth to be defined, but in order to test hypotheses about linearity, at least one more occasion is required. Identification rules become more complicated as the model departs from a basic growth curve. Fewer repeated measures may be required if some parameters are further constrained. Note that the measurement occasions may adhere to a strict schedule that is the same for all cases, or they may be individually varying (see Sterba, 2014).
2. *Were a sufficient number of occasions chosen to adequately estimate the mean trend?* Even if the model is identified, $k + 1$ repeated measures are only barely enough to estimate parameters for a model with k basis curves. More than the bare minimum is generally preferred. Three occasions are required to test a hypothesis of linearity, but six occasions provide a far superior test of linearity. A large number of repeated measures becomes particularly important as the complexity of the hypothesized trend increases.
3. *Does the study span an interval sufficiently long to capture the process?* It is important to ensure that enough time elapses over the course of the study to adequately capture the process of interest. For example, if growth in the height of elementary school children is measured on seven occasions spanning five weeks, clearly not enough time has elapsed to observe meaningful growth, and error of measurement likely will eclipse actual growth. If the trend under study is actually S-shaped (e.g., a learning curve) but data are collected during only a brief window of this process, the trend may appear to be linear. In such cases, a linear LGM might provide spuriously good fit, whereas a more appropriate nonlinear LGM may appear to overfit the data. By the same token, extending measurement beyond what is necessary to capture a trend can waste resources. If a simple linear trend is assessed across 18 occasions, the researcher may be better served to collect data from more cases at fewer occasions.

The key points regarding the scope of the study are that the number and range of repeated measurements need to be considered and justified on the basis of theory and minimum identification requirements. The number of repeated measurements should exceed the minimum, but should not be so numerous as to be wasteful. The interval spanned by the first and last measurement must be sufficient to permit the process under study to unfold.

6. Repeatedly Measured Variables

As in any application of SEM, all measured variables should be defined clearly, or else references should be provided. Some basic requirements of the repeatedly measured variables are that they be reliable and valid, and they should represent some attribute or characteristic that is able to change in

level over time, but not *meaning*. The repeatedly measured Y should not represent stable attributes for which there is no theory of systematic change. In addition, it is necessary to provide a theoretical rationale for expecting not only change, but change of a particular form in a particular direction. If intercept and slope factors are expected to vary across individuals, this expectation should have its roots in theory. In short, it should be realistic and theoretically appropriate to expect change in the chosen Y variable, and Y should be demonstrably reliable and valid.

As in other applications of SEM, special estimation procedures are required for variables that are ordinal, binary, or censored. Application of standard LGM to discontinuous or non-normally distributed variables violates key assumptions necessary for legitimate statistical inference, and thus require special approaches as well.

7. Control Variables

Often researchers may want to examine the effects of some variables on others after controlling for covariates. Chapter 33 in this volume outlines some rules that should be followed in including covariates in SEM, and these rules largely apply in LGM as well. In LGM, two kinds of covariates are distinguished based on their location in the model. *Time-varying covariates* (TVCs), as their name implies, are predictor variables modeled at the level of the repeated measurements. They may be included in the model either as distinct variables predicting Y at each occasion, or as additional loading columns in the Λ matrix in the same way that the variable *time* is included in the Λ matrix. The first method is traditional practice in applications of LGM. The second is identical to how TVCs are included in hierarchical linear models (see Chapter 22, this volume). The LGM framework is flexible enough to permit either specification.

Time-invariant covariates (TICs), on the other hand, are included at the subject level. TICs are exogenous predictor variables used to predict individual differences in aspects of change. For example, in Figure 13.3 we might introduce gender as a predictor of the Intercept and Slope factors. For both kinds of covariate (TVC or TIC), interest may lie in controlling for the effects of covariates, that is, removing them from consideration so that effects of more substantive interest may be interpreted more “purely.” Or, interest may lie in interpreting the effects of the covariates directly.

8. Sampling Method and Sample Size

The requirements regarding sampling method and sample size stated in Chapter 33 still hold when the model in question is a latent growth curve. If stratified or cluster sampling is used, it is crucial that this be considered in the modeling stage by employing design-based estimation methods that take the sampling design into account (such as linearization and replication) or the model-based approach of multilevel SEM; additionally, if disproportionate selection rates were used, inclusion of sampling weights may be appropriate.

It is also essential that studies reporting latent growth curve analyses explicitly address the issue of sample size. As in non-LGM applications of SEM, there are several things to consider when choosing a sample size. First, N needs to be large enough to support the estimation of potentially many free model parameters. Maximum likelihood (ML) estimation is a large-sample technique, and alternative estimation algorithms may require sample sizes much larger than ML. Second, the sample must be large enough to achieve adequate power for rejecting poor models by some criterion of fit. The criterion most commonly used for this purpose is RMSEA. Applications of LGM tend to have high power, but this does not release the researcher from the obligation to demonstrate that power is adequate in a given application. Third, the sample must be large enough for parameters of interest to have small standard errors (and thus narrow confidence intervals and high power).

This is a very important consideration in LGM, where interpretation of parameters is of central interest. Finally, longitudinal studies are typically characterized by missing data due to attrition, death, late entry into the study, and other causes. The total sample size must be large enough to accommodate the amount of missing data. Although missing data techniques such as imputation may be used to fill the gaps in a data set to permit easy analysis, they cannot create information lost to attrition.

Ideally, the researcher should not only meet the minimum sample size suggested by these considerations, but exceed it by a considerable margin. The sample size may be just large enough to exceed the minimum required to achieve adequate power for tests of individual parameters, yet still fall short of the N required to yield usefully narrow confidence intervals. If the sample size is beyond the researcher's control, such as when samples of convenience or archival data are used for analysis, the researcher should still demonstrate that N is large enough to support estimation of a growth curve model and valid interpretation and testing of parameters. The important point here is the sample size should be justified on reasoned grounds, not simply reported.

9. Missing Data and Outliers

Few studies have complete data on all variables for all cases. In long-term longitudinal studies, where the same individuals are followed over time, there are particularly many reasons some observations may be missing for some cases, and missing data in turn may threaten the generalizability of results. These reasons may include attrition due to illness, incarceration, death, data management errors, late entry into the study, lack of subject motivation to follow through, and cancellation of research funding. Late entry and attrition are particularly dangerous in LGM studies, as data missing due to late entry or attrition are typically *not* missing at random. Longitudinal studies with nontrivial amounts of missing data at the beginning or end of the studied trajectory are subject to potentially severe bias in intercept and slope mean and variance estimates.

Four broad strategies for dealing with missing data may be identified: *prevention, deletion, full-information, and imputation*. The best strategy to address missing data is to preemptively prevent data loss by design. Researchers should make every reasonable effort to minimize the proportion of missing data. Data deletion strategies (i.e., pairwise and listwise deletion) use only those cases with complete data for all or some of the variables, and are usually to be avoided. Full-information strategies involve estimating model parameters using all available information, even if some of that information comes from cases with incomplete data. Usable information can be gleaned even from cases with a single valid data point. Imputation strategies use information from existing data about the relationships among variables, and then fill in, or *impute*, reasonable values for the missing data. Complete-case methods are then applied to the imputed data set. Mean imputation, in which missing values of a variable are replaced with the sample mean, should be avoided because it often results in a distribution with unrealistically many cases at the mean value. If imputation is used, multiple imputation (in which several data sets are imputed and the results are averaged across imputations) usually gives the best results. There are other, less often used strategies but these account for the majority of them. Full-information and multiple imputation methods are those most often recommended by methodologists to study missing data. Pairwise and listwise deletion and mean imputation should not be used without extraordinarily compelling reasons.

In LGM, data may be missing by design. For example, in research that combines multiple overlapping cohorts, one cohort may be measured at occasions 1, 2, 3, and 4, while "missing" the fifth occasion of measurement, whereas another may be measured at occasions 2, 3, 4, and 5, while "missing" the first occasion. Combined, all five occasions are represented, but the first and fifth occasions are not as well represented as the other three. This kind of data collection strategy can save time, but it obliges the

researcher to assume that it is legitimate to combine cohorts to form a single trajectory. This assumption can and should be tested rather than assumed. If multiple cohorts are included, data missing by design from one cohort should not be imputed because it can lead to the creation of extrapolated data that are unrealistically congruent with those from other cohorts.

In summary, some general guidelines for reviewing studies with missing data can be developed. The amount and kind of missing data should be explicitly addressed, as should the likely reasons missing data were missing, the steps taken to address missingness in the analysis, and the likely impact of missing data on statistical analyses, the study's conclusions, and generalizability. Full-information and multiple imputation strategies are usually the best choices for addressing missing data. Regardless of the strategy chosen to address missing data, the researcher should justify that choice. It is helpful to report the percentage of missing data for each variable at each occasion. Reporting only the percentage of complete cases does not provide enough information.

10. Software and Estimation Method

For key software considerations the reader is referred to Chapter 33 in this volume. Not all SEM software packages can fit growth curve models. Because the key parameters in LGM include the means of latent variables, the software must be capable of modeling means. Currently, the major SEM software packages capable of employing LGM are AMOS, EQS, lavaan, LISREL, Mplus, OpenMx, SAS PROC TCALIS, and Stata. Because SEM software is regularly updated and improved, it is important to list the name and version of the software used to fit models and obtain parameter estimates.

11. Problems with Convergence, Estimates, and Identification

All of the advice in the corresponding section of Chapter 33 applies to LGM. Errors of identification, convergence, and estimation occur routinely in specifying and fitting models. Accurate documentation of these problems, and the steps taken to remedy them, is essential in reporting results.

It takes a fair amount of skill to properly specify a standard structural equation model, but LGM often requires even greater facility with software and knowledge of the mathematics behind the model. It is easy to make mistakes in specifying a model, and often these mistakes go unnoticed because the software is incapable of distinguishing sensible models from nonsensical ones, or because software will cheerfully provide reasonable-looking results despite serious problems. Here three examples of problems that sometimes occur in the application of LGM, and which may go unnoticed by inexperienced researchers, are described.

First, if a model is underidentified, some SEM software applications (e.g., LISREL and Mplus) may automatically and unobtrusively add constraints to some parameters to render the model identified. Occasionally the researcher is unaware that this automatic identification occurs, and parameters that should not be interpreted are interpreted nevertheless.

Second, negative or boundary values of variance parameters may be reported. Latent growth curve models are fairly robust to estimation problems, but if the model is severely misspecified some very strange things may happen. For example, residual variances may sometimes be estimated as negative (or, depending on the software, constrained to a boundary value of zero). This kind of result is a serious and common estimation error known as a *Heywood case*, and is usually indicative of model misspecification or, sometimes, a sample that is too small. If a residual variance of zero is reported, it is likely a Heywood case rather than a true zero. This kind of error may go unnoticed and unreported because the parameters of central interest in LGM are those related to the intercept and slope factors, not typically the residuals. Thus, if residual variances are not reported, they should be.

Third, covariances among aspects of change (intercept, slopes) may correspond to correlations that lie outside of the logical bounds of -1.0 and $+1.0$. This problem may not be immediately recognizable if only variances and covariances are reported. For example, both covariance matrices below, labeled as Ψ_A and Ψ_B , may appear to be legitimate covariance matrices at first glance, but only Ψ_A is acceptable or “proper.” The covariance in Ψ_A equates to a reasonable correlation of $.28/(.19^{1/2} \times .64^{1/2}) = .80$, but the covariance in Ψ_B equates to an impossible correlation of $.38/(.19^{1/2} \times .64^{1/2}) = 1.09$.

$$\Psi_A = \begin{bmatrix} .19 & .28 \\ .28 & .64 \end{bmatrix} \quad \Psi_B = \begin{bmatrix} .19 & .38 \\ .38 & .64 \end{bmatrix}$$

If an improper matrix is reported, this indicates that an undetected estimation error probably has occurred. Such errors sometimes can be addressed by correcting coding errors, removing outliers, or providing better starting values for the estimation procedure. Sometimes the problem cannot be solved, which indicates that specified growth model may not be appropriate for the data.

12. Data Display and Accessibility

Fitting latent growth curve models requires access either to raw data or to a covariance matrix and mean vector. In line with recommendations endorsed in Chapter 33 of this volume, the data should be reported or made available to permit other researchers to verify or reexamine reported results. Whenever possible—given the constraints of the study, journal space, and proprietary issues—summary information in the form of a covariance or correlation matrix, mean vector, and standard deviations for complete data typically are sufficient to permit reanalysis. If some data are missing or if journal space does not permit reporting summary data, authors should provide instructions informing readers how they may obtain the data.

13. Data-Model Fit

A primary advantage of SEM is that it permits the assessment of fit between the model and data. The advice offered in Chapter 33 of this volume is reiterated here. The χ^2 statistic by itself has only limited usefulness as a fit index because of its sensitivity to sample size and trivial departures from perfect fit. It is usually wise to report multiple (at least three) fit indices drawing from the three broad types (*absolute indices*, *parsimonious indices*, and *incremental indices*). SRMR, RMSEA, and TLI (NNFI) are perhaps the best representatives of these types, but individual researchers may choose not to limit themselves to these, or to include all of them. If RMSEA is reported, its associated 90% confidence interval should also be reported and interpreted. Models may also be compared on the basis of relative fit.

14. Appropriate Null Model

Incremental fit indices like those discussed in Chapter 33 (NFI, NNFI/TLI, and CFI) constitute an important type of fit indices. They express the fit of a substantive model as falling somewhere between the fit of a highly restrictive “null” model and that of a saturated (perfectly fitting) model. The appropriate null model must satisfy two requirements: (a) it must be nested within the most

restrictive substantive model to be tested and (b) it must constrain all covariances among observed variables to zero (Widaman & Thompson, 2003). The null model typically used in SEM is one in which only the means and variances of the observed variables are estimated, constraining the covariances to zero. In ordinary applications of SEM, this null model is perfectly appropriate. LGM, however, requires a different null model. Specifically, if the model to be tested is any growth curve model in which the residual variances are all freely estimated, the appropriate null model is an intercept-only model in which the only free parameters are the intercept mean and the free residual variances—a model with $p + 1$ free parameters that implicitly constrains the variables' means to equality but permits them to have different variances, and constrains their covariances to zero. On the other hand, if the model to be tested is any growth curve model in which the residual variances are constrained to equality, then the appropriate null model is an intercept-only model in which the only free parameters are the intercept mean and the constrained-equal residual variance (only two parameters). Thus, the incremental fit indices reported by default by all SEM programs are incorrect for latent growth curve models, and in fact yield an overly positive evaluation of model fit. They must be computed by hand by explicitly fitting the appropriate null model, obtaining the resulting χ^2 statistic, and manually computing the desired incremental fit index.

15. Model Comparisons

Investigating alternative models of change is a sound and recommended approach to theory evaluation. Compared to the evaluation of individual models in isolation, the comparison of rival models of change has a better chance of eliminating some theories from consideration. For recommendations regarding model comparison, the reader is referred to Chapter 33, Desideratum 14. Model comparison proceeds exactly the same way in LGM as in the more general SEM. Examples of the kinds of models one might wish to compare in LGM could include linear vs. quadratic models, or models in which residual variances are constrained to equality vs. freely estimated.

16. Model Respecification

Latent growth curve models are notoriously poor-fitting by traditional criteria. This poor fit arises not because LGM is unrealistically restrictive, but rather because most other applications of SEM have relatively many free parameters and show unrealistically *good* fit. The trajectories specified in LGM are highly constrained and are not likely to arise by chance in nature. Like any model, growth curve models are merely approximations to reality, and cannot be expected to fit perfectly. However, because tradition and publication pressures have made good fit a necessary component of publishing applications of SEM, the researcher may be tempted to counter instances of poor fit by freeing parameters identified by modification indices (see Chapter 33, Desideratum 15) and fitting the modified model to the same data, resulting in better fit. This temptation should be resisted. A good rule of thumb is that a model may be modified to any degree on the basis of modification indices, but (a) the modifications must be theoretically appropriate and (b) the modified model should be fit to new data to avoid the possibility of capitalizing on chance characteristics of the sample.

Using modification indices is especially discouraged in LGM because relaxing constraints on the model can severely compromise the interpretation of the model as a specific trajectory. For example, if the model in Figure 13.2 were fitted to data and the software reports a large modification index for the loading connecting Y_4 to the slope factor, freeing the loading may improve fit, but the resulting model can no longer be interpreted as a linear growth curve. Furthermore, the model no longer represents a test of the original, theoretically prescribed hypothesis. Even when not

fit to new data, modified models nevertheless have use as descriptive tools. The main point of this desideratum is that modified models may be useful, but tests of such models should not be treated as confirmatory or as strict tests of hypotheses about growth over time.

17. Parameter Estimates and Significance

The end product of model-fitting is a collection of parameter estimates and fit indices. Assuming that model fit is reasonable and that no convergence, estimation, or identification problems persist, it is important to report the magnitude and significance of all model parameters—not only those that are of central interest, but *all* of them. In typical applications of LGM this number is not large. In the basic linear LGM with homoscedastic residuals, for example, there are only six parameters to report: the mean intercept and slope, intercept and slope variances, their covariance, and a common residual variance. More complicated models result in more parameter estimates to report.

A simple way to report parameter estimates is to place the point estimates in the appropriate locations in a path diagram (see Desideratum 4) along with some indication of significance or precision, such as confidence intervals, standard errors, *p*-values, or a system of asterisks indicating levels of significance. The method by which significance is determined is an important and often overlooked consideration. For some parameters—path coefficients and latent variable means—the traditional *z*-tests (dividing the point estimate by its standard error) usually are appropriate. However, for variance and covariance parameters, *z*-tests are not appropriate because such tests require the assumption that the tested parameter is normally distributed over repeated sampling, an untenable assumption for variances and covariances. A better test for variances and covariances is the *likelihood ratio test* (see Chapter 33, Desideratum 14), in which the fit of a model that fixes the parameter to zero is compared to the fit of a model in which the parameter is freely estimated. Depending on the parameter's role in the model, this test may not always be possible or appropriate. Alternatives are to obtain bootstrap confidence intervals for parameter estimates (available in several SEM software applications), likelihood-based confidence intervals (available in OpenMx), or Bayesian credible intervals (available in Mplus).

18. Interpretive Language

Several important points from Chapter 33 regarding interpretive language are reiterated here. First, latent growth curve models, and the theories of change they represent, are never literally “true,” nor can they ever be confirmed empirically. Such statements are extremely misleading and should be discouraged at every opportunity. Models represent formal hypotheses, and hypotheses may fail to be rejected for many reasons, among them low statistical power. In fact, any structural equation model (LGM included) can be made to fit perfectly by freeing a sufficient number of parameters, but perfect fit does not imply a confirmed model. The outcome variable *Y* is not observed as a continuous function of time, so there is no foolproof way to confirm that *Y* values corresponding to unobserved values of time would similarly conform to the trend followed by the observed values, even if perfect fit is observed in the sample. Similarly, growth curve models with zero or negative degrees of freedom would fit *any* data equally well (i.e., perfectly), regardless of what process generated the data. In neither case can perfect fit be taken as support for the researcher's theory of growth. Even identified models with good fit can be described only as tenable in light of the data.

Second, causal language should be used sparingly if at all. A theory's predictions may be causal in nature, but a model's results can never completely support a conclusion that a process is causal, regardless of how well-designed the study may be. Alternative explanations can always be devised. However, confidence that a process is causal may be strengthened by experimental manipulation

(e.g., treatment vs. control), temporal precedence (causes must always precede effects in time), and strong theory that prohibits or limits plausible alternative explanations (e.g., kindergarten may “cause” growth in verbal knowledge, but never the reverse). In applications of LGM, the data are naturally longitudinal, but this does not grant a license to use causal language with impunity. At most, findings may be supportive of a causal process, but can never definitively demonstrate causality.

References

- Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: John Wiley & Sons.
- Byrne, B. M., & Crombie, G. (2003). Modeling and testing change: An introduction to the latent growth curve model. *Understanding Statistics*, 2, 177–203.
- Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods*, 1, 421–483.
- Curran, P. J. (2000). A latent curve framework for the study of developmental trajectories in adolescent substance use. In J. S. Rose, L. Chassin, C. C. Presson, & S. J. Sherman (Eds.), *Multivariate applications in substance use research* (pp. 1–42). Mahwah, NJ: Lawrence Erlbaum Associates.
- Curran, P. J., & Hussong, A. M. (2003). The use of latent trajectory models in psychopathology research. *Journal of Abnormal Psychology*, 112, 526–544.
- Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd ed.). Mahwah, NJ: Erlbaum.
- Hancock, G. R., Harring, J. R., & Lawrence, F. R. (2013). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 309–341). Charlotte, NC: Information Age Publishing, Inc.
- Little, T. D. (2013). *Longitudinal structural equation modeling*. New York: Guilford Press.
- McArdle, J. J. (2012). Latent curve modeling of longitudinal growth data. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 547–570). New York: The Guilford Press.
- McArdle, J. J., & Nesselroade, J. R. (2014). *Longitudinal data analysis using structural equation models*. Washington, DC: American Psychological Association.
- Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: Sage.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.
- Sterba, S. K. (2014). Fitting nonlinear latent growth curve models with individually varying time points. *Structural Equation Modeling*, 21, 630–647.
- Sterba, S. K., & Bauer, D. J. (2010a). Matching method with theory in person-oriented developmental psychopathology research. *Development and Psychopathology*, 22, 239–254.
- Sterba, S. K., & Bauer, D. J. (2010b). Statistically evaluating person-oriented principles revisited. *Development and Psychopathology*, 22, 287–294.
- Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, 8, 16–37.
- Willett, J. B., & Sayer, A. G. (1996). Cross-domain analyses of change over time: Combining growth modeling and covariance structure analysis. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 125–157). Mahwah, NJ: Erlbaum.

14

Latent Transition Analysis

David Rindskopf

Social science theories refer to characteristics (and associated measured variables) that are either continuous (or nearly so) or categorical. For categorical characteristics, we may then distinguish among theories that refer to measurement at a single time point (static) or measurements at two or more time points (dynamic). For dynamic theories, interest centers on the initial distribution of people across categories, and how people transition from a category at each time point to a category (either the same or different) at another time point. The measurements at different time points might be of the same characteristic or of different characteristics (e.g., how does personality type measured at an early age relate to whether a person ends up in a white-collar or blue-collar job later). To complicate the situation further, observed measurements may be thought to be imperfect indicators of an underlying (latent) characteristic. We may now define *latent transition analysis* (LTA) as a statistical model in which (i) latent categorical constructs are defined at two or more time points, (ii) parameters are included that assess initial status and transition probabilities from time i to time $i + 1$ (for most models; for others, we predict further into the future), and (iii) observed variables are imperfect indicators of the hypothesized latent variables. As a simple example, we might theorize that at each age, children either can or cannot conserve volume (in accordance with Piaget's theory). We could give a three-item test to each of a large number of children at ages 3, 4, 5, 6, and 7, and then see whether the data are consistent with the theory, and if so, assess the rate at which children transition from being non-conservers to conservers at each age. The distinguishing characteristics of this study are that (i) the model allows children to be in one of two true states at each time, (ii) the observed test items are categorical (scored right/wrong), and (iii) each child is measured several times (here, over a four-year period).

The roots of LTA are in (i) latent class analysis (see Chapter 12, this volume), conceptually originated by Lazarsfeld (1950a, 1950b, 1959) and systematically developed by Goodman (1974a, 1974b) and Haberman (1974, 1977), and (ii) panel analysis, developed originally by Lazarsfeld and expanded on by Wiggins (1973). The ideas of LTA then grew in several independent but closely related strands, sometimes under different names. General references include Collins and Flaherty (2002), Collins and Lanza (2010), Collins and Wugalter (1992), and Lanza, Flaherty, and Collins (2003). More advanced works include Böckenholt (2005), Humphreys and Janson (2000), Langeheine and van de Pol (1994), Molenberghs and Verbeke (2005), Mooijaart (1998), van de Pol and Langeheine (1990), and Vermunt, Langeheine, and Böckenholt (1999). Example applications include: Chung, Park, and Lanza (2005), who studied substance use among females as they went through puberty

Table 14.1 Desiderata for Latent Transition Analysis.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. A theoretical structure is presented that hypothesizes:	
(i) discrete latent variables occurring at more than one time/age;	
(ii) conditional/predictive relations of earlier to later times/ages;	
(iii) discrete observed measures of latent variables.	I
2. Explicit consideration is given to plausible alternative structures.	I
3. Diagrams of models, if useful for communication of structures, or comparisons among them, are included.	I
4. Equations representing the model(s) are included. These should include allowance for all complications such as multiple groups, nesting (multilevel structures), complex sampling, and so on.	I
5. Parameter identification (unique estimation) is proved or demonstrated.	I
6. A rationale is provided for any restrictions (e.g., equality constraints) used to make model(s) identified.	I
7. Software used to estimate parameters is described.	M
8. Fit statistics used to evaluate model(s) are described.	M
9. A tabular presentation is included of model fit statistics for each model tested, and (where feasible) cell frequencies are provided for possible reanalysis.	R
10. Tabular or graphical presentation of parameter estimates and standard errors (where appropriate) are provided.	R
11. Models retained (as plausible) and rejected (as implausible) are discussed, and a rationale based on fit statistics is provided.	D
12. Any implausible or unusual results in any models not rejected on statistical basis (fit statistics) are discussed.	D
13. Parameter estimates for each retained model are discussed.	D
14. Implications of each retained model are discussed, and a comparison among them is provided.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

(ages 12–15); Graham, Collins, Wugalter, Chung, and Hansen (1991), who studied alcohol and tobacco use among adolescents and the effects of a substance abuse prevention program on that use; and Reboussin, Reboussin, Liang, and Anthony (1998), who studied health risk behavior (in this case, carrying weapons) of schoolchildren over a five-year period.

Computer programs available for estimation of latent transition models include WinLTA (Collins, Lanza, Schafer, & Flaherty, 2002), LEM (Vermunt, 1997), and Panmark (van de Pol, Langeheine, & De Jong, 1996), all of which are either free or inexpensive. Commercial packages such as Mplus (Muthén & Muthén, 1998–2018), Latent Gold (Vermunt & Magidson, 2000, 2003, 2005a, 2005b, 2015), SAS PROC LTA, and LCA (Lanza & Collins, 2008; Lanza, Collins, Lemmon, & Schafer, 2007; Lemmon, Bray, & Chung, 2007) are also available.

Table 14.1 lists desirable characteristics of studies that use LTA. These are discussed in more detail in the corresponding sections of this chapter.

1. Theoretical Structure

The theoretical structure delineation begins with a consideration of the possible discrete states in which a person might be at each time of measurement. For a single construct measured repeatedly,

these states (and the nature of the latent variable) will be the same at each time point. For example, developmental psychologists might want to know whether a child has reached the stage where s/he can conserve volume, and might measure each child at three ages. In other cases, the constructs might be different at different ages: A reading theory might indicate that a child who develops certain word decoding skills by the time of entry into kindergarten will be more likely to be able to read aloud fluently by the end of second grade.

Part of the theory might concern possible versus impossible transitions from one age to another. Without illness or injury, one might expect that a child who can conserve volume at one age will not lose that skill at a later age, so the transition from a more advanced to a less advanced stage would be constrained to have probability zero.

It is the authors' responsibility to make sure that all assumptions are explicitly discussed. These include the number of latent classes, which observed variables measure which latent variables (if there is more than one latent variable), and any restrictions on parameter estimates.

2. Plausible Alternative Structures

As reasonable as our favorite theory might seem, one must always entertain the possibility that it is wrong in some minor (or even major) aspect. Therefore, alternative plausible theories should be explicitly considered. The ideas of "multiple working hypotheses" and "strong inference" should be in every empirical researcher's repertoire; see Chamberlain (1890) and Platt (1964) for a detailed explanation and rationale.

Some alternative structures will also be in the form of latent transition models, and should be tested as described here. Others might have a different structure; most commonly this will be due to some variables (either latent or observed) being continuous or count instead of discrete with a few categories. The researcher should therefore explicitly discuss all plausible alternative structures for the data.

3. Model Diagrams

Diagrams of models can often indicate the main qualitative features in a more easily comprehensible fashion than can equations, although equations will contain the full quantitative specification of the model. Diagrams are also useful when comparing features of two or more models. For latent transition analysis there are no standard methods of representing models, but two methods are generally used. One such method uses the conventions of path analysis models; in this type of diagram there is no explicit differentiation between latent and observed variables. Curved lines with arrowheads at both ends are used to represent unexplained relations among variables; straight lines with an arrowhead on one end are used to represent hypothesized causal inferences.

The other common method is similar to structural equation model diagrams (see Chapter 33 this volume); in these diagrams, circles are used to represent latent variables, and rectangles are used to represent observed variables. The other aspects of path diagrams (curved and straight lines) are retained in these diagrams. When feasible, a diagram should be presented for each distinct type of model tested.

4. Equations Representing the Model(s)

The simplest latent transition models can generally be represented notationally by three categories of parameters. First, there are the (unconditional) probabilities of being in each category of the latent measure at the first time point. For example, at the first time of measurement, children

might be conservers or non-conservers of volume; the probability of being in each category must be estimated. Next are the transition probabilities, which are the probabilities of being in particular states at each future time point given the state at the previous time points. Quite often each state is hypothesized to depend only on the immediately preceding state, which results in a *latent Markov model*. Finally, parameters are needed to account for the probability of being in each category of each observed variable as a function of a person's state in the latent variable underlying the observed variable. Sometimes these latter probabilities are considered to characterize the measurement properties of the observed variables, much as factor loadings indicate how well continuous observed variables measure factors (see Chapter 8, this volume).

In some cases, LTA models are written in terms of equations related to logistic regression or log-linear modeling (see Chapters 16 and 17, this volume). These models are then translated back after estimation into probabilities. Although a model written in the logistic form is perfectly proper and valid, it is often not easily understood even by specialists, let alone researchers in subject matter areas who are not methodologists. For this reason, it is desirable to translate results of applications of LTA into probabilities (either in text, tables, or figures) when feasible.

The basic LTA model can be extended in various ways. The most frequent such extension is to multiple populations. For example, one might theorize that males and females either have different distributions among latent statuses at the first time of measurement, or that they have different measurement properties or transition probabilities. More than one such characteristic may be included in a model (e.g., gender, race, and SES, along with possible interactions), and should be when there are strong theoretical grounds for doing so. LTA models can also accommodate continuous predictors of latent states and transition probabilities.

As with many models, complications can include nesting of people within groups, requiring multi level models or design-based corrections to parameter standard errors. Complex sampling designs may have been used, which also requires care in analysis. Authors should thoroughly describe the design and analysis implications of these situations when they arise in their research.

5. Identifiability of Model Parameters

Some statistical techniques never (or rarely) encounter problems in estimating parameters. For example, multiple regression weights are always estimable if there are enough data, and if predictors are not overly (multi)collinear. Models with latent variables, however, are not the same; sometimes parameters are not estimable no matter how many subjects one has in a sample. Such parameters are not *identified*, to use technical terminology. A non-statistical example might give the general idea: Consider the equation $x + y = 10$. One cannot solve uniquely for x or y ; there are too many unknowns and not enough (independent) equations.

The rules for determining whether each parameter in a model is identified, which would result in the whole model being identified, are not simple to apply. Luckily, there are numerical techniques that in most cases will discover for any particular data set whether each parameter is estimable. Also, many special cases have already been explored, and if one stays within these cases one always knows that (in principle) the model is identified.

Some models that are not identified can be made so by imposing restrictions on parameters. For example, if the same latent variable is measured at several times by the same observed variable, it is sometimes reasonable to presume that the measurement properties of the observed variable do not change over time. In this case, one would impose restrictions of equality on the conditional probabilities of responses at each time. By imposing restrictions, one is estimating fewer parameters, and in many cases this will be enough to make identified an otherwise unidentified model.

Another common type of restriction occurs when there are several times of measurement, and the conditional probabilities of changing from time i to time $i + 1$ are restricted to be the same for

all transition times. This type of restriction is less often theoretically justifiable than measurement restrictions, although if one can argue that a process has reached a steady state then there is more hope for it to be true.

Because most researchers will not be able to prove algebraically that a model is identified, they will have to rely on software to establish empirical identification. They should examine the standard error of each parameter; if any is suspiciously large (under most circumstances, bigger than 3 for the log-linear version of a model), they should suspect problems. If this occurs, and if any observed frequency is zero, one can add .5 (i.e., half a person) to *all* cells and retest for identifiability. If all standard errors now look reasonable, the model can be considered identified.

6. Restrictions/Constraints Used for Identification

Just because one can impose restrictions to make a model identified does not mean that it is right to do so. In the case discussed in Desideratum 5, one might not be justified in assuming that an observed measure is equally good at every age. For example, at lower ages there might be more chance of misunderstandings by the child, which results in greater likelihood of errors. Such possibilities should always be considered, rather than just accepting model restrictions merely because they achieve identification. At the same time, if such restrictions are plausible (or of theoretical interest) they should be tested. Presuming both restricted and unrestricted versions of the model are over-identified, one can compare the fit of the models with and without restrictions of probabilities across groups or times of measurement.

In other cases, restrictions might be more plausible. For example, in comparing males and females, it will frequently be the case that conditional response probabilities to items will be equal for both groups, but unconditional probabilities of class membership or for transition probabilities may differ for the two groups. One can frequently test this assumption by relaxing restrictions one item at a time, searching for what is called *differential item functioning* (DIF) in the psychometrics literature. In the end, it is incumbent upon the authors to justify whatever restrictions or constraints were used to achieve identification of the model(s) being investigated.

7. Software

Although each computer program should provide the same estimates, not all programs will use the same terminology and structure. Authors of research manuscripts should describe the program and its use in a manner that aids those not familiar with it to fully understand what the program calculates. The major differences among programs are in the notation they use. Not only will most substantive researchers be unfamiliar with the concepts of latent transition analysis (which should be explained in the Introduction and Methods sections), but even more so they will be unfamiliar with notation used by different researchers who have developed these models. Therefore one should thoroughly explain the notation that was chosen. As mentioned in Desideratum 4, additional differences might occur in how the model is parameterized (in terms of probabilities or in terms of logistic or log-linear models).

8. Fit Statistics

Fit statistics are used for two main purposes. The first is to test whether a model is in reasonable agreement with the observed data patterns. This is the typical application, which is an extension of the usual chi-square tests of independence in two-way tables with which most researchers are familiar. The problem with such an extension is that if there are a large number of observed variables, the number of people in some (or possibly most) cells of the cross-tabulation will be so small that the

usual test statistics will not follow a chi-square distribution. Some programs contain procedures for constructing bootstrap tests that circumvent this problem.

A second use of fit statistics is to compare the fit of different models. When one model is a special case of another (obtained by imposing restrictions on the more general model), then the likelihood-ratio fit statistics may be subtracted, and referred to a chi-square distribution with degrees of freedom equal to the number of restrictions made. This test has a chi-square distribution even though neither of its constituent fit statistics does. Although it would appear that models varying only in the number of latent classes would be nested (e.g., the two-class model is a special case of the three-class model), the comparison of such models is not straightforward, as the difference in likelihood-ratio statistics is not a chi-square distribution. Research on the comparison of such models is discussed in Nylund, Asparouhov, and Muthén (2007).

When models are not nested, one may compare any of a number of fit statistics that are adjusted for model complexity. These include the Akaike information criterion (AIC), Bayesian (or Schwarz) information criterion (BIC), and other modifications of these. For these statistics, like the chi-square statistics, smaller values indicate better fit. But unlike chi-square statistics, which always decrease as more parameters are added, the AIC and BIC can increase for more complex models due to the penalty that is added.

Often the AIC will favor more complex models than the BIC, and one should inspect not only the overall fit statistic (to see if the BIC is penalizing complexity too much, or the AIC too little), but also the parameter estimates. Sometimes it will be obvious that a model may seem to fit well by one of these criteria, but not be easily interpretable (due to odd parameter estimates), and therefore can be rejected.

One must also take into account that sample size can influence some fit statistics; all unadjusted goodness-of-fit statistics (Pearson, likelihood ratio, and Cressie–Read) will become large when sample sizes are large, even when a model is incorrect in only a minor way. Like all statistical tests, increasing sample size will result in high power to detect small differences from a null hypothesis (in this case, the hypothesis that a model fits the data in the population). The penalty for a BIC is also a function of sample size, and therefore heavily penalizes the addition of parameters (removal of degrees of freedom); this is why it tends to favor simpler models. As much as we would prefer to have absolute standards for model selection, it remains an art: One must consider the sample size, the fit statistics, the adjusted fit statistics, and the parameter estimates for plausible models. With small sample sizes, more than one model may remain plausible, and all such plausible models should be discussed.

9. Tables of Model Fit Statistics

When several models are tested, a table is the most useful way to display the fit statistics for each model. Degrees of freedom for each model should always be included, and if the sample size is large enough (and the observed table has few enough cells) one should include a *p*-value for each model. Similar rules apply here as for the usual cross-tabulation; if all expected cell frequencies are greater than five, there is no problem. If most are bigger than one or two, then the chi-square distribution should apply reasonably well. Some programs will print out various forms of the statistic (Pearson, likelihood ratio, and Cressie–Read); if all are reasonably similar, then most likely the fit statistic and associated *p*-value can be taken seriously.

10. Retained and Rejected Models

Referring to the table of fit statistics, the author should discuss why some models are retained as plausible, and other models are rejected as implausible. Here, overall fit statistics and comparisons

of these statistics are used; later, more refined (though idiosyncratic) decisions can be made on the basis of specific aspects of a model (see Desideratum 12).

When reasonable to do so, one might use a combination of judgments about the fit of each model in isolation and the statistical comparison of fits of sets of models. Sometimes only non-statistical comparisons are reasonable due to small expected frequencies. As in logistic regression, with large cross-tabulated tables many small cell frequencies occur, in which case the usual fit statistics do not have a chi-square distribution even when the model is correct. Comparisons of nested models are still valid under these conditions. Using only p -values from fit statistics (without comparisons) is never wise; one may find that Model A (just barely) fits according to some criterion, and Model B (just barely) does not, but there is no way to decide statistically that Model A fits better than Model B without a direct comparison. One can make qualitative comparisons of AIC and BIC statistics, even for non-nested models, but there is no statistical test comparing the AIC or BIC values for different models.

11. Tabular/Graphical Presentation

Typically only a small number of models fit the data (sometimes we are lucky just to get one!); in this case either tabular or graphical display of the results will be reasonable. Standard errors for all model parameters are usually produced, along with z -statistics (parameter estimates divided by standard errors), by most software. But one must be cautious with these: If standard errors are large, and the parameter estimates are also large, never take the ratio seriously because it will be wrong. Unlike linear models (including factor analysis and structural equation models), non-linear models such as logistic regression (of which latent transition models can be considered a special case) can have parameters that go to boundaries. If the parameters are probabilities, zero and one are the boundaries, and the parameter estimates do not have a normal distribution (so standard errors do not tell us anything useful); if the parameters are on the logistic scale, probabilities of zero and one (or other, more complicated, effects) can go to infinity, and again standard errors are not useful.

12. Implausible or Unusual Model Results

If there is reason to believe that an observed measure is not perfect (and in most cases this is true), then one should suspect any results that show no error of measurement, even if the overall model fit is excellent. Similarly, if one has good reason to believe that no one should go from state A at time i to state B at time $i + 1$, but a large proportion are estimated to do so (e.g., children who can conserve volume at time 1 but cannot at time 2), then one should suspect the veracity of the model. Another indication of problems is if two constructs (e.g., measures at two adjacent times) should be strongly related, and instead they are only weakly related. In the end, there are only a few good general rules here. Knowing which results are plausible and which are not is mostly a matter of knowing (or thinking you know) the subject matter and the model characteristics very well.

13. Parameter Estimates

Even sophisticated quantitative models are often best summarized in simple qualitative terms. For example:

The observed measures are quite accurate reflections of the latent variables, as indicated by conditional probabilities greater than .8 for each item that a person in a mastery latent class will answer that item correctly, and that a person in a non-mastery latent class will answer the item incorrectly.

One need not discuss each individual parameter (which, in any case, would not be a summary). Instead make broad general statements and list exceptions to these generalities.

In most models, two general aspects are important. First, how well do the observed variables measure the latent classes? Second, what are the relations in the latent structure? This would include initial probabilities of being in each class, as well as transitional probabilities from each time point to the next. In models that have multiple groups, a comparison of the parameter estimates across groups should be made for those parameters that are not constrained equal across groups. These estimates should make sense in terms of the original theory from which the model was derived.

14. Implications and Comparisons of Retained Models

If there is more than one model that is consistent with the data, authors should discuss the ways in which their implications are similar, and the ways in which they are different. In some cases, it will be apparent what data would be necessary to distinguish the plausible models (refer again to Platt, 1964); if so, next steps for research that would help resolve remaining issues should be suggested. In some cases, larger sample sizes may be necessary to distinguish the fit of alternative models. In other cases, perhaps more times or different spacings of measurement points may be needed to elicit different predictions of competing models. Another possibility is that more measurements are needed at some (or all) time points.

References

- Böckenholt, U. (2005). A latent Markov model for the analysis of longitudinal data collected in continuous time: States, durations, and transitions. *Psychological Methods*, 10, 65–83.
- Chamberlain, T. C. (1890). The method of multiple working hypotheses. *Science*, 15, 92–96 (reprinted in 1965: *Science*, 148, 754–759).
- Chung, H., Park, Y., & Lanza, S. T. (2005). Latent transition analysis with covariates: Pubertal timing and substance use behaviors in adolescent females. *Statistics in Medicine*, 24, 2895–2910.
- Collins, L. M., & Flaherty, B. P. (2002). Latent class models for longitudinal data. In A. L. McCutcheon & J. A. Hagenaars (Eds.) *Applied latent class analysis* (pp. 287–303). Cambridge: Cambridge University Press.
- Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New York: Wiley.
- Collins, L. M., Lanza, S. T., Schafer, J. L., & Flaherty, B. P. (2002). *WinLTA user's guide version 3.0*. State College, PA: The Pennsylvania State University, The Methodology Center.
- Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, 27, 131–157.
- Goodman, L. A. (1974a). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215–231.
- Goodman, L. A. (1974b). The analysis of systems of qualitative variables when some of the variables are unobservable: A modified latent structure approach. *American Journal of Sociology*, 79, 1179–1259.
- Graham, J. W., Collins, L. M., Wugalter, S. E., Chung, N. K., & Hansen, W. B. (1991). Modeling transitions in latent stage-sequential processes: A substance use prevention example. *Journal of Consulting and Clinical Psychology*, 59, 48–57.
- Haberman, S. J. (1974). Log-linear models for frequency tables derived by indirect observation: Maximum likelihood equations. *Annals of Statistics*, 2, 911–924.
- Haberman, S. J. (1977). Product models for frequency tables involving indirect observation. *Annals of Statistics*, 5, 1124–1147.
- Humphreys, K., & Janson, H. (2000). Latent transition analysis with covariates, nonresponse, summary statistics and diagnostics. *Multivariate Behavioral Research*, 35, 89–118.
- Langeheine, R., & van de Pol, F. (1994). Discrete-time mixed Markov latent class models. In A. Dale & R. B. Davies (Eds.), *Analyzing social and political change: A casebook of methods* (pp. 171–197). London: Sage.
- Lanza, S. T., & Collins, L. M. (2008). A new SAS procedure for latent transition analysis: Transitions in dating and sexual risk behavior. *Developmental Psychology*, 44, 446–456.
- Lanza, S. T., Collins, L. M., Lemmon, D., & Schafer, J. L. (2007). PROC LCA: A SAS procedure for latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 14, 671–694.
- Lanza, S., Flaherty, B. P., & Collins, L. M. (2003). Latent class and latent transition analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 663–685). Hoboken, NJ: Wiley.
- Lazarsfeld, P. F. (1950a). The logical and mathematical foundations of latent structure analysis. In S. A. Stouffer, L. Guttman, E. Suchman, P. F. Lazarsfeld, & J. Clausen (Eds), *The American soldier: Studies in social psychology in World War II, volume IV: Measurement and prediction* (pp. 362–412). Princeton, NJ: Princeton University Press.

- Lazarsfeld, P. F. (1950b). The interpretation and computation of some latent structures. In S. A. Stouffer, L. Guttman, E. Suchman, P. F. Lazarsfeld, & J. Clausen (Eds.), *The American soldier: Studies in social psychology in World War II, volume IV: Measurement and prediction* (pp. 413–472). Princeton, NJ: Princeton University Press.
- Lazarsfeld, P. F. (1959). Latent structure analysis. In S. Koch (Ed.), *Psychology: A study of science, vol. 3* (pp. 476–543). New York: McGraw Hill.
- Lemmon, D. R., Bray, B. C., & Chung, H. (2007). PROC LTA: Latent transition analysis for the SAS System. Paper presented at the 15th Annual Meeting of the Society for Prevention Research, Washington, DC.
- Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. New York: Springer-Verlag.
- Mooijaart, A. (1998). Log-linear and Markov modeling of categorical longitudinal data. In C. C. J. H. Bijleveld & T. van der Kamp (Eds.), *Longitudinal data analysis: Designs, models, and methods* (pp. 318–370). Newbury Park: Sage.
- Muthén, L. K., & Muthén, B. O. (1998–2018). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, 14, 535–569.
- Platt, J. R. (1964). Strong inference. *Science*, 146, 347–353.
- Reboussin, B. A., Reboussin, D. M., Liang, K. Y., & Anthony, J. C. (1998). Latent transition modeling of progression of health-risk behavior. *Multivariate Behavioral Research*, 33, 457–478.
- Van de Pol, F., & Langeheine, R. (1990). Mixed Markov latent class models. In C. C. Clogg (Ed.), *Sociological methodology* (pp. 213–247). Oxford: Blackwell.
- Van de Pol, F., Langeheine, R., & De Jong, W. (1996). *PanMark 3 user manual; PANel analysis using MARKov chains*. Voorburg, The Netherlands: Central Bureau of Statistics.
- Vermunt, J. K. (1997). *LEM 1.0: A general program for the analysis of categorical data*. Tilburg: Tilburg University. Retrieved from www.uvt.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html.
- Vermunt, J. K., Langeheine, R., & Böckenholt, U. (1999). Latent Markov models with time-constant and time varying-covariates. *Journal of Educational and Behavioral Statistics*, 24, 178–205.
- Vermunt, J. K., & Magidson, J. (2000). *Latent GOLD user's manual*. Boston, MA: Statistical Innovations.
- Vermunt, J. K., & Magidson, J. (2003). *Addendum to Latent GOLD user's guide: Upgrade for version 3.0*. Boston, MA: Statistical Innovations.
- Vermunt, J. K., & Magidson, J. (2005a). *Technical guide for latent GOLD 4.0: Basic and advanced*. Belmont, MA: Statistical Innovations.
- Vermunt, J. K., & Magidson, J. (2005b). *Latent GOLD 4.0 user's guide*. Belmont, MA: Statistical Innovations.
- Vermunt, J. K., & Magidson, J. (2015). *Upgrade Manual for Latent GOLD 5.1*. Belmont, MA: Statistical Innovations.
- Wiggins, L. M. (1973). *Panel analysis*. Amsterdam: Elsevier.

15

Latent Variable Mixture Models

Gitta Lubke

Latent variable mixture models serve to investigate heterogeneous populations consisting of two or more clusters of subjects. For instance, a population may consist of subjects who master a study topic and are therefore prepared to take an exam, and subjects who are not well-prepared. Similarly, in a longitudinal setting, there may be groups of subjects who differ with respect to their developmental trajectories. The observed data from a potentially heterogeneous population are modeled using a mixture distribution rather than a single distribution. A mixture distribution is a weighted sum of component distributions, and each of the component distributions is usually assumed to correspond to a cluster of subjects. Model parameters can be specific for each component distribution, and can therefore be used to model differences between the clusters.

Due to the complexity of latent variable mixture models, researchers need to make a large number of decisions during the specification of a model. Decisions include those regarding distributional assumptions, class-specific vs. class invariant parameters, and/or estimating factor variances within class rather than constraining variances to zero. These choices should ideally be theory-driven, but in practice they are often related to practical considerations such as model convergence. In order to permit an evaluation of the quality of a given mixture analysis it is essential that a manuscript covering a mixture analysis includes a description of the decisions as well as the arguments that support the choices. Such a description places considerable emphasis on the methodological details of a given analysis, however, it is a mandatory aspect of using complex statistical methods for the analysis of empirical data.

Literature describing basic latent variable mixture models include Arminger, Stein, and Wittenberg (1999), Bartholomew and Knott (1999), Dolan and van der Maas (1998), Heinen (1996), Jedidi, Jagpal, and DeSarbo (1997), McCutcheon (1987), Muthén and Shedden (1999), Vermunt and Magidson (2003), and Yung (1997). The current chapter focuses on general guidelines for conducting and evaluating latent variable mixture modeling analyses. Details related to analyses using specific types of mixture models can be obtained by consulting the corresponding literature.

It should be noted that mixture modeling is an evolving area. Estimation methods now include Bayesian estimation, and existing models are still being extended. The table of desiderata is based on the current state of knowledge, and summarizes the different issues that researchers and reviewers should consider in a substantive study using mixture models. Details are provided in the subsequent explications.

Table 15.1 Desiderata for Latent Variable Mixture Models.

<i>Desiderata</i>	<i>Manuscript Section(s)*</i>
1. A conceptual description of the models is given and their theoretical underpinnings are explained.	I
2. Information regarding the exploratory and/or the more confirmatory parts of the used models is provided, as are the general assumptions.	I
3. The description of observed measures emphasizes item (or subscale) properties.	M
4. Measurement invariance, and class-specific vs. class-invariant parameters.	M
5. The analysis plan includes a detailed description and justification of all fitted models.	M
6. Model equations are provided and at least briefly explained. Path diagrams may be included to illustrate within-class models.	M
7. Fitting a series of increasingly complex models is practical when the objective is to assess more complex mixture models.	M
8. Explanation is provided as to how and why covariates and class-predicted outcomes (if any) are integrated into the fitted models as proposed. Potential effects of covariate inclusion on class enumeration are mentioned.	M
9. A priori expectations of the power to distinguish between alternative models are described, taking into account expected class proportions, class separation, sample size, and differences among models in the number of estimated parameters.	M
10. Sufficient sets of random starting values are used to afford the replication of the likelihood of the accepted solution with different sets.	M
11. Model fit assessment includes tests to decide on the number of classes, information criteria (ICs), and a discussion of relevant parameter estimates.	M/R
12. Assumptions regarding missing data, and how it might affect estimation of class proportions and within class parameters, should be clearly addressed.	M/R
13. For competing models, results are summarized in tables showing likelihood values, number of estimated parameters, ICs, relevant parameter estimates, and their standard errors.	R
14. If applicable, a justification of additional post-hoc models is provided.	R
15. Model selection is taken into account when drawing inference regarding significance of model parameters.	R
16. If possible, an attempt to validate the class structure using external criteria is provided to strengthen confidence in inference regarding the results.	R
17. Post hoc tests focusing on comparing classes with respect to covariates not included in the main analysis should be carried out properly in a 3-step procedure to account for class assignment uncertainty.	R
18. Input files and information on how to access the data (if possible) are provided.	R
19. The interpretation of results should acknowledge that the number of classes may not reflect the number of distinct groups in the population, and that the selection of a set of adequately fitting model is subject to sampling fluctuation.	D
20. A section detailing the limitations includes a discussion of specific alternative explanations of the results and the potential of lack of power to distinguish between models.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

1. Conceptual Description of the Models

Following the usual outline of the theoretical context of a study, the Introduction of a paper using a latent variable mixture model (LVMM) should provide a clear description how the particular research questions translate into the set mixture models that the researcher proposes to fit to the data. LVMMs permit the specification of a wide variety of submodels that can be related to quite different conceptual interpretations. For example, traditional latent class models (see Chapter 12, this volume) are based on the assumption that the latent classes represent typologies and that observations of members of a given class only vary with respect to random error. Models that specify continuous latent factors within classes permit structural variation, which in turn can be used (for instance) to model severity differences of a disorder within a class, or differences in initial status in a growth model. Due to the conceptual differences of different LVMMs, it is helpful if the Introduction of a paper using LVMM provides clear arguments supporting the choices regarding the type of mixture models used in a particular analysis.

2. Exploratory and Confirmatory Aspects of the Model(s)

Mixture models require specification of the number of latent classes as well as specification of the within-class model structures. Mixture analyses of empirical data commonly compare models with an increasing number of classes, and are therefore exploratory regarding the cluster structure of the sample. In addition, the within-class structure (or aspects thereof) may be unknown *a priori*. For instance, from an exploratory perspective, in a growth mixture analysis it might be part of an investigation to evaluate the necessity of including quadratic effects to model the shape of the growth trajectories, and/or to assess measurement invariance over time or across classes. In more confirmatory applications, questionnaires with a known factor structure may be used, and choices regarding the measurement part of the mixture model are more theory-driven. The Introduction of a manuscript describing a mixture analysis should indicate which aspects of the model have a more exploratory or a more confirmatory character, and which aspects of the model are based on assumptions or on prior research.

3. Description of Measures and Measurement Properties of Items

The basic part of any mixture model used in behavioral research is the measurement of the behavior of interest. The observed measures are usually indicators of traits and/or typologies that are not directly observable. As in any other latent variable analysis, distributional assumptions regarding the observed measures should be clearly described, and ordered categorical data should be handled appropriately. Apart from the distributional assumptions, it is important to realize that the psychometric properties of the scales have a direct impact on the results of a mixture analysis (Lubke & Miller, 2015). The basic requirements regarding the measurement model for latent variables within mixture models are similar to those in other latent variable models. In order to adequately cover the construct of interest, factors should have more than the minimum number of indicators necessary to identify the model, and it is advantageous to use highly reliable items. In the mixture context, it is important to realize that item means or thresholds are directly related to class detection. This is illustrated in Figure 15.1.

Figure 15.1a shows a mixture distribution of a latent factor or trait with two components. The dotted curves correspond to a set of items that measure the trait, and each of the individual curves depicts the probability of answering that item correctly. The set of items in panel A covers the whole range of the underlying trait, that is, the items are located such that there is variability in item scores *within* each class. Figure 15.1b shows two items in the overlap between the two component distributions.

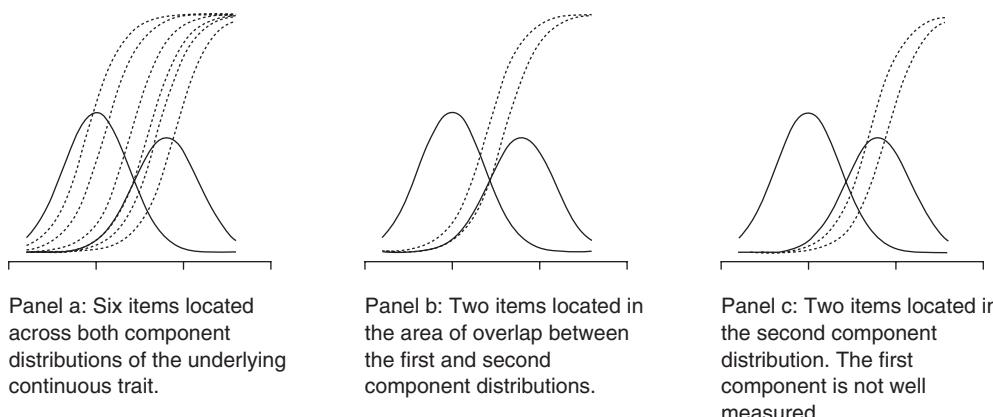


Figure 15.1 Probability of Answering Items Correctly Conditional on a Trait or Factor Following a Mixture Distribution.

These items may be especially important for the estimation of class proportions. Figure 15.1c shows items that are located in the second class. This situation might be representative of a scale, questionnaire, or behavior checklist that is designed to distinguish between more or less affected individuals (e.g., checklists used to identify individuals with psychiatric disorders). The depicted items primarily discriminate between the affected individuals. If such a scale is used in a sample drawn from the general population, then unaffected individuals will score zero on these items, and estimation of variance parameters in the unaffected class is going to be problematic. Conversely, if item locations cover mainly the part of the distribution of the unaffected majority, then the detection of differences in the affected class may be problematic.

In sum, when describing the measures used in a particular analysis it is important to mention for which type of population the scales were originally designed, and whether it can be expected that the items actually measure the full range of the trait in the population that is investigated.

4. Measurement Invariance, and Class-Specific vs. Class-Invariant Parameters

One aspect that deserves special consideration is the choice of class-specific and class-invariant parameters. In the absence of prior knowledge, it would be desirable to fit models with the majority of parameters specified to be class-specific, especially given that incorrectly imposing equality constraints across classes may result in accepting models with too many classes. However, there is a bit of a catch-22. It might be necessary to impose constraints in addition to those that are mandatory to identify the model simply because there are practical limitations to estimating class-specific parameters. Models without any equality constraints across classes on loadings, factor (co)variances, residual variances, or thresholds might not converge, or might result in unacceptable parameter estimates.

The choice of the additional constraints may be theory-driven, or based on the results of the step-wise model building described above. It is important to realize that most constraints correspond to some level of measurement invariance, and have direct conceptual implications. Consequently, it is essential to clarify the decisions made in any given analysis.

As an example, consider the following set of models fitted to observed variables that are multivariate normally distributed within classes.¹ Although it might be subject to discussion whether to start a series of models with the most lenient or the most restrictive model, in the context of mixture models the practical limitations mentioned above often do not leave much choice but to

start with constrained models. An example of a highly constrained model is the latent profile class model,² which has one of the most simple within-class structures. The within-class model specification reflects the assumption that observed continuous variables are uncorrelated within classes (e.g., local independence). This means that non-zero covariation of the observed variables in the total sample (i.e., in the unweighted pooled covariance matrix) is entirely due to mean differences between the classes. On a conceptual level, this is equivalent to assuming that there is a trait underlying the observed variables, but that subjects *within* a class do not differ (have zero variance) on that trait. The model parameters of the latent class model are the residual variances, the observed variable means within each class, and the class proportions. Although this model is usually unproblematic and very fast to estimate, it is a highly constrained model, and might be too crude of an approximation in the cases when subjects within a cluster actually vary on the trait.

A smoother approximation of the true data generating process in the population can be obtained by allowing non-zero factor variances within classes. One can choose to specify a full factor model within classes, in which case one has to decide whether factor variances, loadings, residual variances, and the means of observed variables are class-specific or class-invariant. Instead of class-specific means of the observed variables, one can also choose to estimate factor mean differences between classes. Following the strategy of fitting increasingly lenient models, one can first fit latent class models, and then proceed to fit models that have class-invariant observed variable means, loadings, and residual variances. Such models account for non-zero factor variances within a class, but also assume that observed variables are measurement- invariant across classes.

To test whether measurement invariance is tenable, one can proceed to free the observed variable means, loadings, and residual variances in a stepwise fashion. Note that it is useful to include models with fewer classes in the comparison given that more lenient models usually require fewer classes. Note also that it is informative to check the change in the parameter estimates across models. The process of successively freeing parameters to be class-specific can stop whenever it can be shown that class-invariance is tenable. In case the process is limited by convergence problems or unacceptable parameter estimates, this should be reported to permit the reader to put the results into an appropriate context. For instance, if fitting a model with class-specific residual variances in addition to class-specific loadings does not converge, then it is apparently not possible to test invariance of residual variances. Results would have to be interpreted taking into account that residual variances were assumed to be invariant. Given the direct impact on the conceptual interpretation, choices regarding class-specific and class-invariant parameters have to be described in detail.

5. The Analysis Plan Includes a Description and Justification of All Fitted Models

When conducting analyses involving the fitting and comparison of multiple models, it is tempting to adopt a sequential approach where models are fitted based on the results of previously fitted models. Such a strategy exemplifies capitalization on chance, because decisions on which models to fit next are based on information provided by the sample data. Such a strategy will result in fitting models that adapt to sample idiosyncrasies, and results will be difficult to replicate in a new sample. To avoid this danger, it is important to design an analysis plan *a priori* that includes all planned models. The models should be clearly described, and justified with respect to the research questions to be addressed in the analysis. The analysis plan should be part of the Method section.

6. Model Equations and Path Diagrams

Although the narrative of a mixture paper should provide a clear description of the models that are fitted to the data, due to the complexity of mixture models it is helpful to include the precise

model equations. One should pay particular attention to the subscripts of the parameters given that subscripts provide the necessary information regarding class specificity or class invariance of the parameters, which has important implications for a model's interpretation, and, consequently, for the translation of the research question(s) into models to be fitted and compared. The parts of the paper describing the different models that are fitted to the data can then be linked to the model equations. The combination of equations and narrative can provide an unequivocal explanation of the analysis that is carried out. In addition to the equations, researchers may choose to include path diagrams depicting the within-class measurement models. Path diagrams provide a quick visual reference and can enhance the readability of the manuscript. As with the equations, path diagrams should clarify which parameters are class-specific and which are class-invariant.

7. Fitting Increasingly Complex Models

Estimating complex models can be problematic for at least two reasons. First, as the complexity increases, there is usually an accompanying increase in the number of potential misspecifications. Misspecifications can lead to convergence problems and/or to biased results. Second, fitting complex models is computationally intensive, and unusable results are especially unpleasant after a time-consuming estimation attempt.

It is therefore practical to include simpler models in the analysis plan that have the advantage of being easier to estimate. For instance, second order growth mixture models (i.e., multivariate growth mixture models) have the advantage that measurement invariance can be tested across time. However, fitting an unconstrained second order growth mixture model as a first model might be impractical. A model where class loading, factor variances and covariances, and residual variances are specified to be class-specific may not converge or may result in unacceptable parameter estimates. Even if the goal is to fit such a model, it is useful to evaluate the measurement model at each time point separately before combining the time points. Fitting a series of increasingly complex models can provide detailed information especially if the most complex models fail to converge.

A similar stepwise approach, consisting of fitting increasingly complex models, can be followed when fitting factor mixture models (FMMs). In the case when questionnaire data are analyzed that serve to indicate multiple factors, it is often useful to carry out an initial exploratory factor analysis (EFA; see Chapter 8, this volume) on a subset of the data that is *not* used for the main analysis. The available set of data is split into an exploratory, smaller part, and a larger part that is used for the main analysis. The exploratory subset can be used for EFA and/or other exploratory data analyses. Note that EFA is based on assuming a single homogeneous population, which is incorrect if the population is in fact clustered. It is therefore important to keep in mind that the EFA results are based on the *unweighted pooled* covariance matrix, that is, covariances between items may in part be due to mean differences between classes and not adequately reflect a factor structure within classes. However, an EFA can still provide some indication of the item properties. Items with very low reliabilities can be detected and removed from the main analysis. In addition, if the EFA shows that the items have a simple structure, one might consider fitting models to the different factors separately as a first step in order to investigate the cluster structure without having to invest considerable computation time trying to fit multifactor models. Especially with ordered categorical observed data, fitting multifactor models is computationally intensive.

A manuscript describing a mixture analysis should outline the analysis strategy and the steps the lead to the specification of the set of models used in the main part of the analysis. Providing the results of initial analyses that are done in a smaller exploratory subset of the data will often enhance the confidence in the final results.

8. Integrating Covariates and Class Predicted Outcomes

In general, the inclusion of covariates with substantive effects is helpful because it increases the separation between classes. As a thought experiment, suppose we have two classes and a binary covariate that perfectly predicts class membership. All subjects scoring, say, zero on the covariate will have zero probability of belonging to one of the classes and a probability of 1 to belong to the other class. This means that the classes are completely separated on the covariate. As a result, the multivariate distance between the latent classes computed for the observed variables and the covariate jointly will be larger than the distance computed for the observed variables only. Increased separation between classes is directly related to improved class assignment and increased power to distinguish between alternative models.

A second advantage of integrating covariates is that the model is embedded in a larger conceptual context. Latent classes can be characterized in reference to covariate effects, which may enhance confidence in the results. The same argument holds for integrating distal outcomes that are predicted by class membership and/or within-class factors.

The selection of covariates should be theory-driven, and included in the analysis plan. Covariates can be specified to have effects on class membership, and/or on factors and observed variables within classes. It is important to realize that omission of direct covariate effects on within-class factors or variables can lead to biased results, which is very similar to the omission of direct effects in measured or latent variable path analysis. Class proportions, as well as even the direction of covariate effects, can be incorrect if important direct effects are omitted. It is therefore necessary to provide arguments supporting the choice of covariate effect incorporated in the set of mixture models. In case of lack of prior knowledge, direct effects should be tested. When reporting the results, it should be mentioned whether covariate effects or effects on distal outcomes are based on a priori expectations or whether integration of the effects is exploratory.

9. Expectations Regarding the Power to Discriminate among Alternative Models

Not all mixture analyses are exploratory and consist of comparisons of models with alternative within-class structures. Most analyses, however, compare models with an increasing number of classes. Model comparisons are usually based on information criteria such as the BIC that penalize for the number of parameters in the fitted model. Model choice therefore depends, in part, on the number of additional parameters that are estimated when fitting a model with an additional class. This is especially important when models are compared that use ordered categorical data. For instance, cases where it is unrealistic to assume that the item thresholds are class-invariant, models with class-specific thresholds are compared with an increasing number of classes. If items have a 5-point Likert response format, for example, then adding a class can imply a substantive increase in the number of parameters, and can have the consequence that a model with fewer classes is selected (Lubke & Neale, 2008). On the other hand, a researcher might want to compare a model with measurement invariance constraints to more lenient models with class-specific measurement parameters. In this case the measurement-invariant model with k classes might actually have fewer parameters than the non-invariant model with $k - 1$ classes (see the empirical example in Lubke & Neale, 2008), making the model with more classes more likely to be selected.

In addition to the number of parameters, sample size is clearly an important factor in class detection and discrimination between alternative models. Unfortunately, it is difficult to provide rules of thumb given that sample size requirements depend on class separation, model complexity, response format, and, as illustrated in Figure 15.1, on item properties. Depending on these factors, analyses

for very simple latent class models may-be carried out probably with as few as 30 subjects, whereas other analyses require thousands of subjects.

Monte Carlo (MC) simulations can provide an indication of the expected power to discriminate between alternative models. A large number of data sets may be generated under a given model, say Model A. Next, Model A and an alternative Model B, for instance a model with one additional class, are fitted to the data sets. Model comparisons are carried out for each pair of Model A and Model B fitted to an individual data set, and then the relative proportions may be determined in which Model A and Model B are preferred. MC options are integrated in some software packages for mixture analyses (e.g., *Mplus*).

An important caveat of MC methods is that their results are based on the fact that at least one of the fitted models (i.e., the data generating model) is a true model. In practice, the set of fitted models is unlikely to contain the true model. The true data generating process in the population underlying most human behavior is obviously very complex, and the data generation for simulations is a crude simplification. Researchers should be aware that model comparisons between two models that are fitted to real data where both models contain various degrees of misspecifications do not necessarily have the same properties as simulated comparisons.

10. Random Starts

It is well-known that the likelihood surface of latent variable mixture models has numerous local maxima. As a result, it is necessary to start the estimation using different sets of random starting values. The number of starting values that is necessary to obtain stable results can depend heavily on the psychometric properties of the items (e.g., reliability, variance, location with respect to the trait), the complexity of the fitted model, and the degree of misspecification. It is useful to start with a lower number of sets of starting values, and check the likelihood and the stability of parameter estimates when increasing the number of sets. Replication of the likelihood using different sets of starting values enhances confidence in the solution; however, replication of the likelihood value is neither a sufficient nor a necessary requirement to ensure that a global rather than a local maximum has been found.

Some software packages first compute a limited number of iterations for a given number of sets of starting values, subsequently select a user-specified number of solutions with the best ranked likelihoods, and then iterate those until a convergence criterion is met. Software packages differ in how starting values can be manipulated by the user, hence it is necessary to clearly mention the software package and version number in addition to how starting values were used.

11. Model Fit Assessment

Model fit should be assessed in the context of the power to detect potentially small classes and to discriminate between alternative within-class models. Model comparisons can be based on information criteria and other indices, such as the bootstrapped likelihood ratio test statistic. In exploratory settings, models with more or less constrained within-class structures should be compared. A difficulty in choosing a small set of “best-fitting models” is related to the fact that that less constrained models (i.e., model with a large number of class-specific parameters) may be incorrectly rejected when compared more constrained models due to lack of power (Lubke & Neale, 2008). Given the fact that issues of power, class detection, and detection of class-specific parameters is highly dependent on the particular data and models under consideration, it may be desirable to refrain from narrowing the choice to a single best-fitting model, and rather present a small set of models that may be equally adequate to describe the structure of the data at hand.

12. Missing Data

Recent research has addressed the effect of missing data on the estimation of class proportions and within-class parameters especially in the context of growth mixture analyses (e.g., Lu et al., 2011). As is the case more generally, great care should be taken to include auxiliary variables in the data collection, and subsequently in the model(s), that can explain drop-out in longitudinal studies, or missingness in cross-sectional studies.

13. Presentation of Results

The standard errors of the parameter estimates provide an indication of the stability of a given fitted model. This is especially true for the estimates of factor and residual variances within classes. Parameter estimates should be reported together with their standard errors.

Results of model comparisons should be summarized in tables. Preferably, one table provides the information criteria together with the number of estimated parameters and the log-likelihood values. In addition, a table showing class proportions and parameter estimates of interest is usually a useful reference.

14. Justification of Additional Models

Based on the results of *a priori* planned models, an analysis is often extended and additional models are fitted to the data. As in any other statistical analysis these additional models should be presented as post hoc exploratory analyses. Additional models can be a very useful source of information, for instance, planned analyses can be extended with additional covariates, or the necessity may arise to fit models with within factor structures that differ from the planned models. However, it is helpful to provide a clear justification why these models are included in the analysis.

15. Model Selection and Parameters Estimates

When presenting model results concerning the selected best-fitting models it is tempting to report parameter significance as provided by most software packages. However, it is important to realize that a significance level for a given parameter reported in software output is conditional on model selection: it does not take into account that several models have been fitted to the data, and a subset of “best-fitting models” has been selected. Using the same data for model selection and for assessing model parameter significance leads to potentially dramatic increases in Type I error (Hurvich & Tsai, 1990; Lubke & Campbell, 2016).

16. Validation

As mentioned in Desideratum 7, effects of class predicting covariates and distal outcomes provide a conceptual context for a given mixture model, and can be useful to validate the class structure. When designing a study, one might therefore consider collecting data on potentially interesting covariates or outcome variables, and generating hypotheses about how latent classes are related to these supporting variables. Validation in a different sample drawn from the same population is of course desirable. Unless the original sample size is extraordinarily large, splitting the sample for validation purposes is discouraged given that sample size plays a crucial role in class detection and discrimination among models with alternative within-class structures. The sensitivity of mixture models to sampling fluctuation has not been thoroughly investigated, and generalizations of a given cluster structure may be highly sample specific. Unless a validation is carried out, interpretation of results has to acknowledge this limitation.

17. Post Hoc Class Comparisons Should Be Interpreted with Caution

Published studies sometimes report post hoc tests using posterior probabilities. Subjects are assigned to their most likely class based on the highest posterior class probability, thus transforming the latent classes into groups, and subsequent group comparisons are carried out with respect to variables that were not included when fitting the mixture model. This type of post hoc class comparisons can be problematic due to the error rates in assigning subjects to classes. This is especially true in the case of unbalanced class size. As illustrated in Figure 15.2, small classes have a high prior probability of incorrect assignment even if classes are relatively well separated. The left panel shows the distribution of a factor in two classes of equal size with a Mahalanobis distance of 1.5. The expected probability of incorrect assignment is symmetric. However, if one class contains only 25% of the subjects, as shown in the right panel, then the a priori probability of incorrect assignment in the small class is substantially higher, namely .5.

In addition to expectations related to class size, it should also be noted that posterior class probabilities are computed using parameter *estimates*, and therefore contain the accumulated uncertainty from those estimates. Studies have shown that assignment error rates can be considerable (Tueller & Lubke, 2010), and studies examining the effect of assignment error on post hoc class comparisons are under way. Given the current uncertainty concerning the validity of post-hoc testing, results should be interpreted with caution.

18. Provide Input Files and Data (if Possible)

It is usually helpful to provide commented input files of at least one of the fitted models in an appendix. If possible, one should provide (a link to) the empirical data so other researchers can repeat the analyses or fit alternative models. In the event that the empirical data may not be made public, a (link to the) full set of parameter estimates should be provided so that other researchers can generate simulated data.

19. Interpretation of Results

The interpretation of any given mixture analysis should be placed into the context of the psychometric properties of the items or scales, sample size, class separation, and response format of the items, since all of these factors influence the power to detect classes, to detect class-specific parameters, and to obtain stable results.

Mixture models are often used to detect qualitatively different clusters of subjects. However, it is important to note that the latent classes correspond to the components of a mixture distribution,

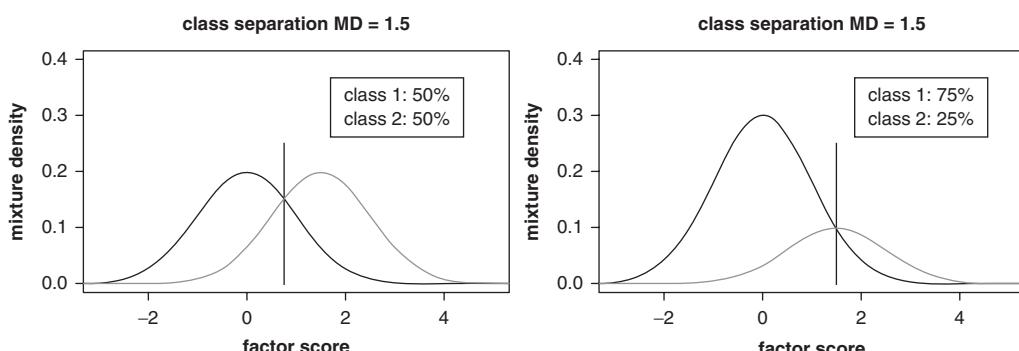


Figure 15.2 The Effect of Class Size on Expected Incorrect Class Assignment.

and that the component distributions are used to approximate the joint distribution of the observed variables. Figure 15.3 illustrates the approximation of a distribution of a single continuous variable that is slightly skewed. The three normal mixture components in the right-hand panel have equal variance. It is easy to visualize that two components with unequal variance might provide a similarly good approximation.

Translating the idea to multivariate observed variables, suppose the observed distribution is multivariately skewed. Depending on the degree of multivariate skewness, a mixture of two or more multivariate normal component distributions can approximate the skewed distribution better than a single multivariate normal distribution. In other words, mixture components correspond to areas under the joint distribution of observed variables that contain subjects with similar response patterns. An interpretation along this conceptualization of latent classes is desirable. Furthermore, research focusing on model selection uncertainty has highlighted the fact that model selection results can fluctuate considerably if the same set of models is fitted to new samples drawn from the same population (Preacher & Merkle, 2012). It is therefore advisable to present the results of a model comparison in a cautious manner, and interpret the similarities and dissimilarities of a small number of best-fitting models rather than limiting the discussion to a single best-fitting model.

20. Limitations of the Study

A number of journals have the standard of requesting a paragraph detailing the limitations of the current study. Such a section should be a staple especially in the context of latent variable mixture models due to the number of assumptions and potentially subjective decisions involved in the model building and model fitting process. It is useful to discuss the reasonableness of meeting model assumptions, the tenability of imposed constraints, and the potential detrimental effects of these decisions. The section on the limitations can also include a discussion of potential alternative interpretations of the current results (e.g., lack of power to detect small classes, overextraction of classes due to skewness).

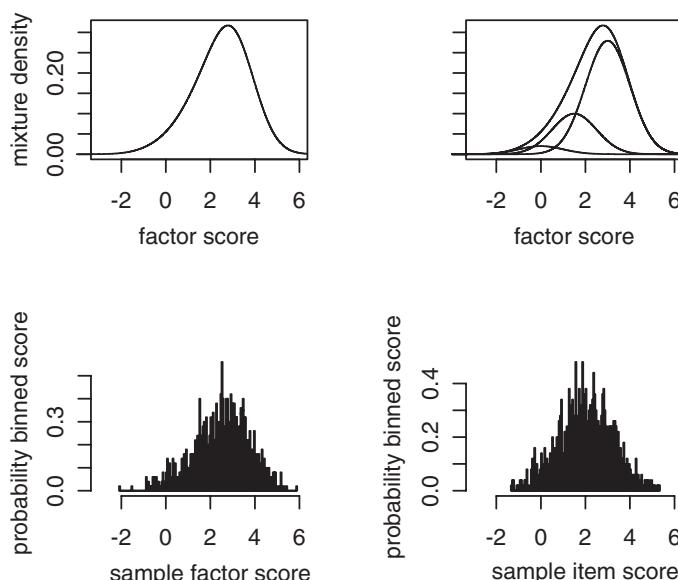


Figure 15.3 Approximation of a Skewed Distribution Using a Mixture of Three Normal Component Distributions with Equal Variance.

Notes

- 1 Measurement invariance for categorical data has been discussed in the multigroup context by, for instance, Muthén and Asparouhov (2002) and Millsap and Tein (2004). Some aspects that are specific to the mixture settings are described by Lubke and Neale (2008).
- 2 Latent profile models are latent class models for continuous outcome variables.

References

- Arminger, G., Stein, P., & Wittenberg, J. (1999). Mixtures of conditional mean- and covariance structure models. *Psychometrika*, 64, 475–494.
- Bartholomew, D. J., & Knott, M. (1999). *Latent variables models and factor analysis* (2nd ed.). London: Arnold.
- Dolan, C. V., & van der Maas, H. L. J. (1998). Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika*, 63, 227–253.
- Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks, CA: Sage.
- Hurvich, C., & Tsai, C.-L. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, 44, 214.
- Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite mixture structural equation models for response based segmentation and unobserved heterogeneity. *Marketing Science*, 16, 39–59.
- Lubke G. H., & Campbell, I. (2016). Inference based on the best-fitting model can contribute to the replication crisis: Assessing model selection uncertainty using a bootstrap approach. *Structural Equation Modeling: A Multidisciplinary Journal*, 23, 479–490.
- Lubke G. H., & Miller, P. (2015). Does nature have joints worth carving? A discussion of taxometrics, model-based clustering, and latent variable mixture modeling. *Psychological Medicine*, 45, 705–715.
- Lubke, G. H., & Neale, M. C. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research*, 43, 592–620.
- McCutcheon, A. L. (1987). *Latent class analysis*. Thousand Oaks, CA: Sage.
- Millsap, R. E., & Tein, J. Y. (2004). Model specification and identification in multiple-group factor analysis of ordered-categorical measures. *Multivariate Behavioral Research*, 39, 479–515.
- Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Mplus Webnote no. 4. Los Angeles, CA: University of California, Los Angeles. Retrieved from www.statmodel.com/download/webnotes/CatMGLong.pdf.
- Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55, 463–469.
- Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods*, 17, 1–14.
- Tueller, S., & Lubke, G. H. (2010). Evaluation of structural equation mixture models: Parameter estimates and correct class assignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 17, 165–192.
- Vermunt, J. K., & Magidson, J. (2003). Latent class models for classification. *Computational Statistics & Data Analysis*, 41, 531–537.
- Yung, Y. F. (1997). Finite mixtures in confirmatory factor analysis models. *Psychometrika*, 62, 297–330.

16

Logistic Regression and Extensions

Ann A. O'Connell and K. Rivet Amico

Logistic regression (LR) is a type of regression analysis in the family of models more broadly known as *generalized linear models* (GLMs). LR provides a versatile and flexible modeling strategy for the analysis of binary data in the form of dichotomous outcomes, typically designated as either $Y = 1$ for success or $Y = 0$ for failure. Extensions to LR exist for ordinal and multinomial data, and LR may also be used with binary data that have been grouped or summarized to represent the proportion of successes across multiple trials. The LR model is used to predict the probability of success, also known as the *response probability*, conditional on one or more predictors. For dichotomous outcomes, and letting \underline{x}_i represent the collection of predictors for the i th person in the sample, we can write this response probability $P(Y = 1 | \underline{x}_i)$ as $\pi(\underline{x}_i)$. LR is also referred to as a logit model because it uses a logit link function to transform these conditional response probabilities into the natural log of their odds, called *logits*, where the odds are a quotient comparing the probability of success to the probability of failure. Thus, $\text{logit}(\pi(\underline{x}_i)) = \ln\left(\frac{\pi(\underline{x}_i)}{1 - \pi(\underline{x}_i)}\right)$. Logits are useful in regression modeling because they form a continuous measure that spans the real line, unlike probability which is bounded between 0 and 1, or the odds, which have a lower bound of zero. The logits serve as the outcome being modeled in logistic regression, and a model's estimated logits can be easily back-transformed into estimated probabilities. Like standard linear regression models for continuous outcomes, LR models use single or multiple predictors that may be categorical or continuous, allow for polynomial terms or interactions between predictors, permit user-driven entry decisions or iterative methods (e.g., forward or stepwise), and provide model fit diagnostics and residual analyses.

Examples of LR are readily found in nearly all research areas in the social sciences, as variables that are appropriately understood or defined as dichotomies exist in virtually any substantive area of study—dropping out of, or persisting in, school; presence or absence of a chronic condition or risk behavior; attendance; pregnancy; incarceration; recidivism; completion of a program; and intervention success versus failure. Examples where ordinal logistic regression is commonly employed include models of ratings on severity of pain (1 = little to no pain, to 4 = severe pain) and analysis of ordinal quality of life scales. Similarly, multinomial logistic regression can be used to model outcomes such as choice of occupation. These examples are similar to the extent that the specific logistic regression models used are intended to estimate one or more underlying success probabilities.

Table 16.1 Desiderata for Logistic Regression and Extensions.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. Measurement of the outcome/response variable as dichotomous, multinomial, or ordinal is described and justified.	I
2. Literature supporting design of the study as well as the proposed model of the outcome/response is reviewed and summarized; model theory is consistent with the purposes of the study and the research questions or study hypotheses.	I
3. Core elements of the logistic regression model (or extensions) are identified: sampling model (distribution of interest); link function (transformation linking predicted values to observed values); structural model (expression of transformed outcome as a function of a set of predictor variables).	M
4. Choice of link function for dichotomies or extensions is described and justified: logit, probit, complementary log-log, multinomial, cumulative logit, etc.	M
5. Sample size is identified and justified; sampling strategy and mode(s) of data collection are fully described.	M
6. Software package is identified; response to be modeled is clarified; weighting methods, if necessary, are described and justified.	M
7. Predictor variables are identified, and their measurement is described and justified; evidence for reliability and validity is provided. Selection process explained/justified.	M
8. Coding of all categorical predictors is fully described.	M
9. Parameter estimation strategy is identified.	M, R
10. Extent of missing data is clearly reported for all variables, and methods for accommodating missing data are described.	M, R
11. Choice for standardization of predictors (none, partial, full) is explicated and justified; limitations to any standardization strategy must be noted.	M, R
12. Overdispersion is investigated and adjusted for as necessary.	M, R
13. Choice for hypothesis testing for variable effects is justified (likelihood ratio tests, Wald's χ^2); if appropriate for the study design, the justification for trimming of variables is provided and the resulting competing models are statistically compared.	R
14. Interpretation of parameter estimates is provided for all variables and their interactions (if included). Graphical strategies may be used to enhance interpretation of interactions and/or products of predictors.	R
15. Model assessment strategies are provided, including investigation of multicollinearity, linearity in the logit, and existence of outliers; assessment of parallel/proportional odds; in situations where separation or complete separation occurs, offending variables are identified and corrective measures undertaken.	R
16. Results of residual diagnostics are reported; impact of outliers and extreme or unusual observations are clarified.	R
17. Multiple summary statistics of model fit are presented. These include results of the Hosmer-Lemeshow test; model deviance; chi-square difference tests for goodness of fit and comparisons of competing models; and pseudo- R^2 values.	R, D
18. Categorical assessment of model fit is provided, via classification tables and associated statistics. Percent of correct predictions cannot be the only criteria for classification accuracy; stronger supporting statistics include, for example, τ_p or λ_p .	R, D
19. Final model presented is credible, addresses the research questions/hypotheses, and is supported through literature and theory. Causal language is not used except when justified through study design.	M, R, D

* I = Introduction, M = Methods, R = Results, D = Discussion.

Hosmer and Lemeshow (2000) is widely regarded as the classic text on LR. We also recommend the texts by Agresti (2013, 2007), Allison (1999), Collett (2003), and McCullagh and Nelder (1989). Three relatively brief but excellent descriptions of applied LR can be found in Harlow (2005), Tabachnick and Fidell (2012), and Wright (1995). More advanced treatment on the GLM can be found in Dobson (2002) and McCulloch and Searle (2001). Suggested resources for extensions of LR to other kinds of discrete or polytomous responses include Clogg and Shihadeh (1994), Harrell (2015), and O'Connell (2006) for ordinal outcomes; Agresti (2013), Fox (2008), and Long (1997) for multinomial outcomes; and Snyder and O'Connell (2008) for discrete time-to-event data.

Discriminant analysis is an additional method historically used for analyzing qualitative (dichotomous, multinomial) outcomes, and was a topic in the first edition of this book (Huberty, 2010). Discriminant analysis is a classification method for which LR can be viewed as a more recent alternative. LR is more versatile than discriminant analysis—particularly when predictors include both continuous and categorical variables—with results that are easier to interpret and estimation methods that are not restricted by multivariate data assumptions (Agresti, 2013; Lei & Koehly, 2003; McLachlan, 2004). Given that research studies in the social sciences typically make use of both quantitative and qualitative predictors, LR is a preferable and more plausible approach in practice.

The major focus in this chapter is on LR for dichotomous outcomes. However, the desiderata below are relevant to extensions of the LR model, and we note specific additional details within the text to accommodate models for ordered and unordered multinomial data.

1. Nature of the Response Variable

The response variable in LR refers to the outcome dichotomy of interest. Discussions regarding the response variable in LR must occur early in the manuscript and address whether the variable's dichotomy is a natural binary phenomenon (e.g., voting in the next election or not) or the result of an artificial dichotomization of an underlying scale or measure (e.g., categorizing children as overweight or not based on a cutoff BMI percentile rank, using median splits, or basing cutoff decisions on a certain number of standard deviations above or below the mean). Creating artificial dichotomies is an approach that is poorly regarded in most areas of inquiry. In fact, a recent examination of this practice upholds what historical reviews have long argued: forced dichotomization yields a crude categorization of a potentially useful continuous measure, subsequent loss of information, and a flawed understanding of the phenomena being studied (MacCallum, Zhang, Preacher, & Rucker, 2002). In terms of the impact of this process on logistic regression models, Taylor, West, and Aiken (2006) documented the loss of statistical power that occurs under different categorizations of a continuous response variable, and found that the loss was actually greatest when the outcome was dichotomized. Thus, from a substantive as well as a statistical perspective, the response variable for logistic regression should be based on a true underlying binary phenomenon. The application of LR to coarsely derived dichotomies needs to be accompanied by a carefully articulated justification and understanding of the limitations and consequences of this practice.

In addition to a qualitative description of the response variable, the Introduction section of manuscripts based on results of LR should contain a discussion of the prevalence or expected prevalence of the target response within the population of interest. This base-rate information is critical for informed understanding of the complex interplay between sample size, the number and measurement of covariates, the expected relations between covariates and response probability, and the use of the model for explanation or classification (or both). While the sample-based distribution of the dichotomy and the independent variables should be provided in the Methods and Results sections (see Desiderata 5, 7, and 15, below), the introduction of research using LR must address the anticipated distribution of events based on an articulated theory or previous research in the content area. As the targeted response probability becomes more extreme (closer to 0 or 1), there is an increased likelihood of the model being affected by numerical problems in the data including

complete or quasi-complete separation or the presence of zero-cells (see Desideratum 15). Both of these numerical problems tend to be readily identifiable by extremely large standard errors for variable effects. Thus, because the conceptual and operational definition of the dichotomous outcome and its anticipated and real distributions given a set of hypothesized predictors have critical implications in research design and the statistical and substantive interpretations of results, these aspects of LR modeling should be clearly addressed beginning in the Introduction section of the manuscript.

2. Support for Model Theory

Models fit through LR should be guided by the same philosophical framework as other regression or prediction models (see Chapter 23, this volume). Thus, the capacity to inform the field and the credibility of the model depends on the theoretical underpinning of the model in relation to the investigated phenomena. A thorough introductory discussion of the LR model will clearly define the theoretical basis of the operationalization of the variables under study and the manner in which the constellation of variables modeled coexist. Moreover, because LR models, like standard regression models, evaluate the effect of predictors relative to other variables specifically included in the model, careful consideration of the literature and theory of the phenomenon in question must also address inclusion criteria for variables in the LR model and why some may be excluded.

Interestingly, LR deviates from multiple regression in a critical way regarding decisions on inclusion or exclusion of variables in the final regression model. In most purely exploratory prediction models, an initial analysis might include all available predictors, targeting for elimination from the model any variables that are not statistically significant, generally at an alpha level of .05 or, more conservatively, at .01. Use of this strategy in LR may lead to erroneous elimination of covariates that confound the relation between a critical risk factor and the outcome. Consequently, elimination of model variables on the basis of their level of significance can lead to flawed interpretations of variable effects, particularly if a confounder is inappropriately removed (Tabachnick & Fidell, 2012).

Hosmer and Lemeshow (2000) detailed a careful model building approach that begins with identification of a collection of theoretically or scientifically relevant variables, and then uses results from univariate analyses between each predictor and the outcome as an initial variable inclusion strategy. They recommended using a liberal level of significance in both univariate and a series of multivariable models to protect against removal or non-identification of confounders. Overall, their argument was for the researcher to use his or her best discretion as to the development of the model and the inclusion of all scientifically relevant explanatory variables. Clearly, these decisions can be made only with strong theoretical support drawn from existing literature or through adequately conceptualized studies framed around a body of previous research. The support developed in the Introduction of a study should make specific reference to the variables anticipated to be relevant in understanding the dichotomous outcome and in answering the research questions or study hypotheses. Later decisions to modify the resulting LR model should be articulated and reasonably justified.

3. Core Elements of the Generalized Linear Model

In a fully informed methods section of a manuscript, three key elements of the LR model—as a generalized linear model—must be clearly stated or be implicit in the models description: the sampling distribution, the link function, and the structural set-up of the model. We discuss each of these for the GLM in general for context and then specifically with respect to LR for dichotomous outcomes.

Generalized linear models have been used to represent the behavior of a wide variety of discrete outcomes in practice, and their theoretical connection with standard linear regression models simplifies their application. This simplicity is evident by considering the general structure of the GLM (McCullagh & Nelder, 1989; Nelder & Wedderburn, 1972), formally characterized through three specific features: (1) a random component describing the anticipated distribution of the response

variable and based on the exponential family (e.g., normal, binomial, Poisson); (2) a linear component describing how a transformation of the expected value of the response variable can be written as a linear predictor based on a given collection of covariates or explanatory variables; and (3) a link function specifying the connection between the original and the transformed responses.

In LR, the binomial distribution is typically used to describe the behavior of binary data or proportions and thus forms the random component. The distribution of a binary random response variable, Y_i , in logistic regression is generally expressed as $Y_i \sim B(1, \pi_i)$, where B indicates the binomial distribution, 1 indicates the number of trials (which equals 1 because each individual forms his or her own trial), and represents the probability of a successful outcome on the i th case or trial. The mean and variance of this Bernoulli random variable are given by π_i and $\pi_i * (1 - \pi_i)$, respectively. Thus, the probability of success is heteroscedastic across cases—for each observation, the variance is different and depends on the expected value.

The linear component of the LR model describes how a transformation of the expected values can be written as a linear function of a set of p predictors (covariates):

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} \quad (1)$$

The corresponding link function, η_i , describes the connection between the observed responses and the transformed responses. If we were to base this connection on the identity link by letting $\eta_i = \pi(\underline{x}_i)$, where $\pi(\underline{x}_i)$ represents the success probability for the i th case given the set of x covariates, the model's predictions could potentially fall well outside the 0,1 range for probability. Further, the errors for binary data are non-normal and heteroscedastic, invalidating this linear probability model on several grounds. Thus, a suitable transformation of the expected values is required in order to construct the linear component of the model and honor the bounded nature of the binary responses. Several transformations are possible, but the logit link is often selected due to the simple and straightforward interpretation of model results in terms of odds and odds ratios (McCullagh & Nelder, 1989).

The logit link function is the natural log of the odds of success, or logit ($\pi(\underline{x}_i)$), where the odds of success is a quotient comparing the probability of success to the probability of failure. Thus, the linear component of the logistic regression model can be written as:

$$\eta_i = \text{logit}(\pi(\underline{x}_i)) = \ln \frac{\pi(\underline{x}_i)}{1 - \pi(\underline{x}_i)} = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip} \quad (2)$$

The logit maps the expected values onto the real line from $-\infty$ to $+\infty$. Within the logistic regression model, each slope (holding all other effects constant) represents the change in the logit that is expected to occur given a one-unit change in the predictor. Exponentiating a regression coefficient (i.e., e^β) yields an odds ratio (OR) that describes the association between that variable and the outcome in terms of odds. An OR of 1.0 implies that the predictor has no associative effect on the odds of positive response. Small values of the OR (< 1.0) indicate that the odds of success tend to decrease as the predictor increases by one unit; larger values of the OR (> 1.0) indicate that the odds of success tend to increase as the predictor increases by one unit. Odds ratios are routinely reported by statistical packages, and represent a measure of association between each predictor and the binary outcome; they are non-negative and range from 0 to ∞ . The ORs can be interpreted directly, or they can be used to calculate a percent change in the odds given a one-unit increase in a predictor based on the following formula: $100\% * [\text{OR} - 1]$.

In LR, the linear model describes how the log of the odds of success varies by the set of predictors. For predictions based on the logit link, the antilog (or inverse) will provide a prediction for the odds conditioned on that set of predictors:

$$\exp(\text{logit}(\pi(\underline{x}_i))) = \frac{\pi(\underline{x}_i)}{1 - \pi(\underline{x}_i)} = \exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}) \quad (3)$$

Using this expression to solve for probability shows how the probability of success tends to vary as an inverse logistic function of the collection of covariates (McCullagh & Nelder, 1989):

$$\pi(\underline{x}_i) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip})} = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}))} \quad (4)$$

Although LR modeling has many nuances in addition to the primary features of GLMs, these three aspects are the basic building blocks for understanding the LR model. Appropriate use and presentation of LR modeling strategies need not make extensive detailed description of the theory involved in GLM, but reference to the key features (random component, linear component, and link function) should be present in a manuscript's Methods section.

4. Declaration of Link Function

The logit link function for dichotomous data yields what is commonly called the *binary logistic regression model*. Given the simple interpretation of this model in terms of odds and probability, it is often a researcher's preferred choice. Another frequently used model for binary data is the *probit regression model*. Probit regression uses a transformation of the probability of success based on the normal cumulative density function rather than log-odds; the inverse of the probit link is the cumulative standard normal distribution. Parameter estimates and standard errors will be different between probit and logit models given that they are based on different transformations of the data, but their overall interpretations of substantive effects will tend to be similar. Given their similarity, choice between logit or probit is often based on a researcher's familiarity with one of the two approaches. Fox (2008) described several advantages of the logit model over the probit model, emphasizing the simplicity of the logit model. However, a pragmatic investigator may want to compare and contrast goodness-of-fit statistics across competing options. A third link option for binary data is the complementary log-log (*clog-log*) transformation of the success probabilities. The inverse of the clog-log function is the extreme value distribution. Unlike the logit or probit transformation, the clog-log link is asymmetrical which becomes advantageous when analyzing discrete time-to-event data. Such data can be considered discrete and dichotomous when for each case in the sample it is known whether an event occurred or did not occur within a specific interval.

Analyses for multinomial or ordinal data utilize these as well as other link options (Agresti, 2013; Fox, 2008; Long, 1997; O'Connell, 2006). When outcomes are not ordered in a meaningful way, *polytomous* (or *multinomial*) logistic regression is typically applied using baseline category logit models using an underlying logit link. Given k outcome categories or qualitative response values, the analysis proceeds as a simultaneous set of $k - 1$ logistic regressions, where one of the k outcome categories serves as the baseline or referent category against which all others are compared. Thus, parameter estimates for the logit for a given category is relative to the same baseline category. Results thus depend on the particular category chosen as baseline. The multinomial logistic regression model (with the last category as the baseline category) can be written as:

$$\ln \frac{\pi_k(\underline{x}_i)}{\pi_K(\underline{x}_i)} = \beta_{0(k)} + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} \quad \text{for } k = 1, \dots, K - 1 \quad (5)$$

An alternative to application of baseline category logits for polytomous data involves treating the data as a series of nested dichotomous, which cohesively form a set of consecutive binary partitions

of the data. For example, an outcome with four categories labeled 1, 2, 3, and 4 could be consecutively partitioned into a dichotomous series as 1 versus 2, 3, and 4; 2 versus 3 and 4, and finally 3 versus 4. The resulting analysis is often referred to as a *continuation ratio (CR) logit model*, with corresponding *CR link*, and is particularly useful when the outcome values are ordered in a meaningful way. Choice of the CR logit link yields a series of conditional logits (and thus, conditional probabilities), that help distinguish among cases that have reached a particular response category but not progressed further from those cases that “continue” to advance to a higher response category. O’Connell (2006) described how the continuation ratio approach, when applied to discrete time-to-event data, corresponds to the discrete proportional hazards model. In fact, based on how the data are set up to construct the underlying dichotomies, the clog-log link and the CR link will provide identical model interpretations. However, ordinal data need not represent discrete times in order to use either link.

A more common link function for ordinal data is the *cumulative* (also known as *proportional odds*) link. The cumulative odds (CO) link function partitions the data sequentially, but utilizes the full set of responses in each binary partition. For example, a four category response variable would have three successive dichotomies: 1 versus 2, 3, and 4; 1 and 2 versus 3 and 4; 1, 2 and 3 versus 4). Parameter estimates thus represent the effect of each predictor in terms of cumulative logits, which in turn can be used to estimate cumulative probabilities of a case being at or below (or, at or above) a particular response category. The cumulative odds model can be written as follows, although we note here that some software packages utilize an alternative parameterization of the model in which the regression coefficients are *subtracted* from the intercept with corresponding adjustments in interpretation (see O’Connell, 2006 for details):

$$\ln \frac{\pi_k(x_i)}{1 - \pi_k(x_i)} = \beta_{0(k)} + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots \beta_p X_{ip} \quad \text{for } k = 1, \dots, K-1 \quad (6)$$

Both the CR and the CO link make assumptions about the behavior of predictor variables across the series of underlying binary models, referred to as the *parallel* and *proportional odds* assumptions, respectively. Essentially, this implies that the effect of a particular predictor remains constant across each underlying binary model. While contributing to a parsimonious model, the plausibility of this assumption should be verified and considered when choosing a link function.

It should be clear from this discussion that many alternatives are available to a researcher working with dichotomous, multinomial, or ordered response variables. For outcomes with more than two categories, all logit-link related options correspond to a particular set of binary partitions of the data, and thus interpretations of variable effects can be made by extension from binary logistic regression (Harrell, 2015; O’Connell, 2006). Yet regardless of the choice for link function, it is incumbent upon the researcher to clearly identify and justify the decision process used in making their choice, because alternative approaches might impact on the substantive interpretation of the data.

5. Sample Details

Parameter estimation for GLMs and thus LR are based on maximum likelihood estimation, a large-sample methodology. Consequently, larger sample sizes are typically required for logistic regression than might be expected based on a standard linear regression. In addition, there are several inter-related factors that can impact the sample size necessary for reliable estimation of model coefficients or detection of an experimental effect. These include the base rate or response probability within the population of interest (rareness of the event), the difference in sample size between the two response categories (success versus failure), the number of observations per covariate pattern (sparseness

of the data), the type of covariates included in the model (continuous versus categorical), and the expected number of events per covariate. Freeware or commercial sample size and power analysis packages are available that can estimate a desired sample size for given effects in LR. However, it is the rare case in which the desired statistical model can be adequately represented by the assumptions imposed by power software. Hosmer and Lemeshow (2000) have reported that the type and number of covariates covered in power programs for LR are often limited, and the situation has not changed very much in recent years. For example, a recent study by Hemmert, Schons, Wieske and Schimmelpfennig (2016) confirmed lack of progress in this area, but also determined that goodness-of-fit estimates for LR and ordinal/multinomial extensions are strongly affected by design features including sample size and the distribution of responses across outcome categories.

As a first step in determining sample size, researchers could use power software or hand-calculations based on straightforward assumptions regarding factors related to sample size, and adjust the results as necessary to mimic the actual design. As an event becomes increasingly unusual or rare within a given population, larger samples must be taken to ensure an adequate capture of sufficient cases to submit to a logistic regression analysis. Hosmer and Lemeshow (2000) and Allison (1999) described how sampling on the dependent variable in LR is protected against the selection bias inherent in standard linear regression models when sampling disproportionately on the outcome. Oversampling can help ensure sufficient numbers of events without biasing the odds ratios, although the intercept needs to be adjusted for the sampling fraction within both outcome groups. For analyses of extremely rare events (say, that occur with probability in the range of .05 or smaller), King and Zeng (2001a, 2001b) described a related oversampling selection adjustment that enhances the data collection process as well as improves understanding of uncommon occurrences.

When classification is the focus of the logistic regression procedure, additional attention must be paid to the relative sizes of the outcome response groups, given that LR tends to place participants into the larger group (Finch & Schneider, 2006). As a result, the misclassification rate for smaller groups can become extremely large when the groups are very unequal in size.

Data sparseness is a problem that frequently plagues LR. The reliability of estimation weakens when sample sizes are small or when there are few observations per similar covariate pattern in the model. Sparse data patterns are particularly likely to occur when the LR model includes continuous covariates, given that there is expected to be a different covariate pattern for each individual in the sample. Evidence of estimation problems due to sparseness of data from small samples or in samples with continuous covariates are reflected in very high standard errors.

Some researchers have recommended minimum sample sizes of at least 50 observations per predictor for logistic regression (Aldrich & Nelson, 1984). Hosmer and Lemeshow (2000) discussed an extension of the minimum sample size criteria suggested by Peduzzi, Concato, Kemper, Holford, and Feinstein (1996) and recommended that for multivariable logistic models the sample size of the smallest response group be at least as large as $10(p + 1)$, where p is the number of predictors in the model. Thus, the sample size of the smaller response group, and not the overall sample size, should be the main consideration in issues pertaining to power, including the number of appropriate parameters for a given model in a given sample or, alternatively, determining sufficient sample sizes to approach *a priori* models of a given complexity.

For readers to fully understand the quality of the research study, the Methods section of a manuscript also must provide information on the mode of data collection (mail survey, face-to-face interview, observation) and the sampling strategy utilized (simple random, cluster, etc.) as addressed in Chapter 35 of this volume. In summary, manuscript authors should always provide sufficient methodological detail to demonstrate the utility of their desired and obtained sample sizes, carefully describing the impact of sample construction on generalizability of results and the impact of sample size on the capacity to detect statistically significant effects.

6. Software and Weighting

Under the assumption that data are drawn using a simple random sample, software packages are relatively comparable in their logistic regression procedures. However, analysis of complex survey data involving stratification, multistage, or cluster sampling often requires software capable of a design-based rather than a model-based approach. A design-based approach incorporates sample design features in the form of weights for each observation's primary sampling unit (PSU), strata, and cluster, into the analysis and adjusts the standard errors of variable effects accordingly. Sampling weights provide protection against potential bias inherent in the unweighted estimator and allow researchers to make valid inferences from the sample to the population the sample was designed to represent. Most commercial software and freeware including R can conduct logistic regression analyses based on complex survey data. There are some differences across software packages in estimation or statistical testing procedures and in how missing data are accommodated; thus, the software and version used should be reported by manuscript authors.

Software packages also may differ in how they internally refer to the dichotomous responses (e.g., 1 versus 2, or 1 versus 0), or in the designation of either outcome as the success response. Given that the odds for the complement of an event is simply the inverse of the odds for that event, the regression weights would retain the same magnitude but the signs would be reversed if the outcome codes were reversed (e.g., for modeling failure rather than modeling success). Thus, the choice has little impact on how the model is interpreted, although the researcher should clearly indicate which outcome is being modeled through the LR.

Finally, LR models based on data from case-control studies, in which the sampling process depends on the response variable, can proceed as if the data were collected through a simple random sample. Case-control studies are generally undertaken to ensure sufficient numbers of a rare event; consequently, event cases are oversampled at a rate much higher than that of controls. While the intercept is affected by sampling on the response variable, the remaining regression coefficients in the logistic model are not. Breslow's (1996) review of case-control studies remains one of the critically important ones in this field and should be read and cited by researchers employing this design. Extensions to the simple case-control design, including complex surveys involving stratification or one-to-one matching, induce several estimation and inferential complexities (Scott & Wild, 2003). Manuscript authors should justify and explain their sampling methods and adjustments in sufficient detail to allow critical examination of the validity of published results.

7. Identification and Measurement of Predictors

As with any type of statistical modeling, careful examination of each predictor or covariate is essential. Identification of potential predictor variables should be predicated primarily on theoretical grounds and supported through the literature review (see Desideratum 2). Procedures for investigating bivariate relations between variables as well as between each potential predictor and the dependent variable should also be described. Decisions to consider interaction terms (products of predictor variables) should be justified based on prior knowledge of relations among variables within the domain being studied. All relevant study variables including background and demographic variables should be clearly described; a summary of these variables for the study sample can be succinctly presented in a table or other summary form within the Methods section of the manuscript.

In LR models, predictors can be continuous, dichotomous, or categorical. The description of classification and coding systems used for dichotomous or categorical predictors must be complete (see Desideratum 8). The measurement of continuous predictors should be clearly articulated in the Methods section, including previous support for validity of the measurement process and careful

evaluation and description of the predictor within the sample data in terms of distribution and reliability. As with most modeling strategies, unreliability in measures of predictor variables introduces error into the model and such measurement errors can attenuate estimated LR slope coefficients (Stefanski, 2000).

In research studies where a large number of variables may be considered for their potential as predictors, the procedures for determining variable selection for inclusion or exclusion during the modeling process need to be described and justified in sufficient detail for other researchers to be fully informed regarding validity of the intended approach and the resulting regression model. Variables may be included *a priori* based on defined or anticipated relations with the outcome, or they may be selected through computer-driven approaches (stepwise, backwards elimination, forward selection, etc.). Reliance on computer-driven approaches, however, does not guarantee a credible final model. In all selection procedures, care must be taken so that inclusion or exclusion decisions are not made solely on the basis of a bivariate relation between each potential predictor and the dependent variable, since assessment of individual contribution of a given predictor within a multivariable regression model is established relative to all other predictors in the model. Including irrelevant predictors that share association with other predictors can inflate the standard errors associated with these regression coefficients, and excluding relevant predictors can lead to cases where “third variable” effects are inappropriately attributed to the predictor set included. Proposed methods for building the LR model must include consideration of complex variable relations, the theoretical justification for each variable’s expected contribution as a potential predictor within a multivariable model, and evidence of strong psychometric properties supporting the measurement of each variable.

8. Coding of Categorical Predictors

Within the Methods section, authors should include a clear description regarding their treatment of all categorical variables. For example, a description of a categorical variable for “education level” of a study participant must identify the specific education levels assigned to each category. In addition, the coding scheme corresponding to each categorical predictor when included in the regression model must be explicit.

Coding schemes are used within all regression-based analyses to completely identify group membership on a given categorical predictor. Depending on the software package used for analysis, there are quite a few options available for the coding of categorical predictors in LR, mirroring similar strategies for inclusion of categorical variables in standard linear regression. Some of the major statistical software packages have default coding schemes along with choices for alternative systems built into their logistic regression procedures that will automatically recode categories of a predictor when prompted by subcommand requests. Different coding schemes can easily be selected and applied to categorical data based on the needs of a particular analysis; alternatively, researchers can choose to do their own coding of categorical data within their data file based on a system that represents critical questions of interest regarding differences between categories. Most often, researcher-constructed schemes involve the creation of a series of dummy (or indicator) codes, effect codes, or a series of orthogonal contrast codes. As with all coding decisions it is incumbent upon the researcher to ensure that the final result for representation of category differences clearly represents the intended questions of interest.

9. Estimation of Parameters

The parameters of interest in most LR models are the regression weights. The method used most often for estimation of model parameters in LR is maximum likelihood (ML). The ML estimates

are the values for the parameters that maximize the likelihood function and thus provide the largest probability of producing the observed data. Mathematically, it is more convenient to maximize the log-likelihood (LL) rather than the likelihood itself; multiplying the LL by -2 creates a quantity called the *deviance* that can be used for hypothesis testing purposes to compare competing models (Hosmer & Lemeshow, 2000). The larger the deviance, the less well the fitted model reproduces the actual data; thus, smaller deviances are preferred. Statistical tests for model comparison are discussed in Desiderata 13 and 17.

ML estimation is based on large-sample theory, and estimates will typically be biased in small samples. Non-convergence of the iterative ML process generally occurs in cases where separation or complete separation is encountered, when zero-cells are present in the data, or when the data are sparse (see Desideratum 15). These problems become more likely when sample sizes are small. Another cause for non-convergence involves overfitting by including more parameters in the model than the data can adequately support. Alternatives to ML include exact methods, which can be applied in small samples or when non-convergence prevents parameter estimation. Hosmer and Lemeshow (2000) described two additional alternatives: iterative weighted least squares, and the use of discriminant function analysis (see Huberty, 2010) which was relied on in early work on logistic regression. However, ML is the method predominantly used today.

In terms of manuscript preparation, the Methods and Results sections should identify the estimation strategy used for determination of parameter estimates. If alternatives to ML estimation are used, the selected procedure should be explained and justified. It may be helpful, but not necessary, for researchers to identify the numerical strategy used during the ML estimation process. Commercial statistics packages differ in the numerical procedures used to iterate and converge on the ML estimates. For example, SAS uses a Fisher Scoring algorithm (equivalent to iterative reweighted least squares), while SPSS uses the Newton-Raphson technique. However, for logit models on binary data, these procedures provide equivalent parameter estimates.

10. Missing Data

Logistic regression is susceptible to the same potential biasing effects of missing data on model covariates as are other statistical models. The default method used in statistical packages for cases with missing data is listwise deletion, affecting the sample size as well as the sample's validity. For the most part, generalizability of statistical results hinges on the match between the characteristics of the obtained sample and those of the population from which it was drawn. Even when samples are drawn at random, missing data interferes with this match—particularly if the missingness follows a non-random pattern. Several excellent sources offer detailed information on analysis options in the presence of missing data; a historical review of missing data procedures and a comprehensive evaluation of approaches can be found in Enders (2010) and Schafer and Graham (2002).

Missing data on an entire case—often referred to as “unit non-response”—is often addressed at the analysis stage through the application of sample-weights for non-response or through post-stratification on known population characteristics such as age or gender distributions. Neither method, however, guarantees that the effects of non-response have been eliminated (Korn & Graubard, 1999). Thus, efforts to eliminate unit non-response should begin at the sample design and data collection stage, rather than at the analysis stage. The strategies taken to limit unit non-response (e.g., repeat call-backs, or the use of proxy-based information) should be clearly described in the Methods section, and if the research calls for sample weighting or post-stratification approaches to compensate for unit non-response, this process also must be articulated in detail within the Methods section. During the presentation of results, the degree of unit non-response (e.g., percent refused, percent unavailable) must be provided, and results of any analytic

procedures (weighting schemes, post-stratification methods) or design-based approaches (percent completed after callbacks, percent completed by proxy) to adjust for unit non-response must be described with sufficient information to allow readers to evaluate the research findings in light of the approaches taken and to anticipate and understand the limitations involved in these strategies.

A second kind of non-response occurs when individual cases are missing responses on one or more items. Item non-response could occur for predictors or for the dependent variable of interest. Schafer and Graham (2002) pointed out that the procedures for dealing with missing dependent variables do not differ significantly from procedures for missing predictors; however, missing dependent variables in a study utilizing LR could seriously affect the patterns observed in the data and consequently our understanding of the phenomenon under study—particularly if the missingness is related to that dependent variable. For dependent variables that might be considered rare or hidden within a population (under-age drinking and driving; thoughts of suicide), care must be taken to prevent missing occurrences as much as possible. Disproportionate sampling (Desideratum 5) may be utilized here to assist in replacing cases with missing outcomes, as well as in attempts to model why certain persons in the sample offer differential response (present versus missing) on the outcome of interest.

There are many imputation methods available for estimating a replacement value or substitute for a missing item, some more reasonable than others. For example, mean-substitution is generally considered one of the weakest forms of imputation and is not recommended (nor is retention and analysis of cases containing only complete information on all variables). The text by Little and Rubin (1987) is likely the definitive resource on the analysis of incomplete data, but the treatment of missing data remains a rich and evolving research area. The current “state of the art” for working with studies involving missing data, according to Schafer and Graham (2002), assumes the data are missing at random (or completely at random) and include maximum likelihood procedures in missing data problems such as the EM algorithm (Dempster, Laird, & Rubin, 1977; Little & Rubin, 1987) or multiple imputation methods. In the case that items are *not* missing at random, sensitivity analyses can be used to gauge the bias that might result under certain prescribed and relevant situations.

Strategies for protecting against or limiting the effects of unit or item non-response need to be clearly described and detailed in the Methods and Results sections of a research manuscript. Sufficient information is required so that readers can evaluate the results and potential biases in model estimates, and anticipate any limitations to the research findings based on the approaches taken by the researcher.

11. Standardized Regression Coefficients

In LR, the raw (or unstandardized) regression coefficient, b_p , for a predictor variable, X_p , can be interpreted as the expected change in the log odds of success given a one-unit change in X_p , controlling for other variables in the model. Similar to linear regression, comparing the absolute size of the raw regression weights across multiple predictors and using the size of these raw weights as a marker for relative influence is not a reasonable practice, particularly when the predictor variables are measured in different scales or metrics. Mirroring concerns regarding the use of standardized regression coefficients in ordinary least squares (Bring, 1994), options for creating and interpreting standardized regression coefficients in LR continues to be a topic of interest among researchers. Specifically, coefficients in LR can be either partially or fully standardized, or be derived based on information theory (Menard, 2002, 2004a, 2004b). Insufficient knowledge or awareness about the estimates produced by different standardization processes can lead to their misuse and inappropriate substantive conclusions. Statistical packages vary in terms of options for standardization

methods. For example, SAS offers a partially standardized regression coefficient, while SPSS no longer reports a standardized result. In any case, if a researcher decides to standardize regression coefficients, sufficient justification for the selected approach as well as a summary of the advantages and disadvantages of that approach for identification of relative importance of a predictor must be adequately presented in the Methods and Results.

An alternative use of the unstandardized coefficients in LR is to interpret them in terms of their odds ratios. Menard (2002) recommended basing substantive results for categorical variables or variables with definitive units of measure (length, cost, counts, etc.) on unstandardized regression coefficients or their corresponding odds ratios, and reserving the use and interpretation of fully standardized LR coefficients for other kinds of variables, such as those measured on a scale like self-concept or attitudes. Likewise, Hosmer and Lemeshow (2000) emphasized interpretation of variable effects in terms of clinical and theoretical importance, rather than in terms of relative contribution to a prediction model. Overall, in situations where researchers report and interpret standardized coefficients, principled reasons supporting the selected standardization method must be described in the Methods and used to clarify interpretation of results.

12. Overdispersion

The LR model for binary response data models the dispersion of the dichotomous outcome, Y , as a binomial random variable, or more simply, as following a Bernoulli distribution, given that the outcome of success or failure is observed on only a single trial for each participant in the study. The general form of the binomial variance is $\sigma^2 = m(\pi)(1 - \pi)$, where π is the probability of success, and m refers to the number of trials. If the response Y is observed on a single trial, $m = 1$ and the mean, π , alone determines the variance. Overdispersion can occur in situations where $m > 1$; these are models for which the response probability for the i th unit is determined by summing the number of successes observed over m trials. When the linear logistic regression model is applied to such proportion or percentage data, the data often exhibit more variability than would be expected based on the binomial variance. This overdispersion is sometimes referred to as *extra-binomial variation*, and its presence suggests heterogeneity in the data that is not accounted for by the model.

Overdispersion is quite common in practice, and can also occur when there is unaccounted for clustering or correlation within the data. Other causes include the presence of extreme observations or outliers, or when the underlying probability of success across each of the m trials is not constant. Unfortunately, contributions to overdispersion cannot easily be distinguished (Collett, 2003). The impact of overdispersion is revealed in standard errors for LR coefficients that are smaller than they should be, leading to increased Type I errors for tests of variable effects and subsequent flawed understanding of the relations being examined.

Dobson (2002) reported that overdispersion may be present if the deviance from a fitted model exceeds its degrees of freedom (generally determined as $J - (p + 1)$, where J is the number of replicated covariate patterns and p is the number of predictors in the model); however, this appraisal assumes that the model is correctly specified and that suspected outliers or other issues have been appropriately addressed. McCullagh and Nelder (1989) described how the covariance matrix for the regression coefficients can be rescaled based on the ratio of the deviance to degrees of freedom. The scaling factor or dispersion parameter, α , reflects the degree to which the model variance should be inflated relative to the binomial variance: $\text{Var}(Y) = \alpha * [m(\pi)(1-\pi)]$. This same strategy can be used for overdispersed multinomial or ordinal models as well.

Tests for the scaling factor are available through statistical software packages, and models can be adjusted by incorporating α as a weighting factor on the variance estimates to compensate for the degrading effects of overdispersion. Alternatively, models that assume a specific form for the

overdispersion may also be applied, such as the beta-binomial model for the analysis of proportions (Collett, 2003; McCullagh & Nelder, 1989). However, McCullagh and Nelder argued against the selection of a statistical model based on a specific assumed form for observed overdispersion on the grounds that mathematical convenience should not take precedence over the scientific plausibility of a chosen model. In support of this argument, they reported that models including an adjustment based on the dispersion factor generally perform better than beta-binomial models.

Overdispersion must be investigated for any logistic regression models of proportions or percents (i.e., when $m > 1$). However, all manuscripts utilizing LR should include at least a brief discussion within the Methods section describing how overdispersion is to be investigated and identifying potential factors that might contribute to potential overdispersion (e.g., sampling design, omitted variables). If scaling is required, the Results section of the manuscript should report the size of the estimated scaling factor, and results should be clearly interpreted in light of any adjustments for overdispersion.

13. Hypothesis Testing for Individual Parameters

For $j = 1$ to p independent variables, the regression weights in the LR model represent the change in the logit for each one-unit increase in X_j , controlling or adjusting for the effects of the other independent variables in the model. The regression weights can be exponentiated to yield the odds ratio for each variable ($OR = \exp(\beta_j)$). Strong associations between independent variables and the outcome typically are represented by ORs further from 1.0, in either direction (see Desideratum 3). Statistical significance of an OR typically is assessed by testing if the regression coefficient, β_j , is statistically different from zero through one of three tests: Wald, score, or likelihood ratio.

In the Wald test, the parameter estimate for the effect of each independent variable in a logistic model is divided by its respective standard error, and the results are squared to represent a value from the chi-square distribution with one degree of freedom under the null hypothesis of no effect. Most major statistical packages report Wald chi-square statistics for each variable in the fitted model. However, Wald statistics can be problematic in small samples, samples with sparse cells, or samples with many covariate data patterns including samples with continuous independent variables. In these situations, statistical tests based on the likelihood function are preferred. The score test for the contribution of an independent variable in the model relies on derivatives of the likelihood function but is not directly available in many statistical packages. However, SPSS does use a score test in stepwise procedures to determine when variables enter or exit a developing model. Finally, the likelihood ratio test is generally regarded as the most reliable test for the contribution of an independent variable to a model, and is based on the difference in deviances between a model which contains that variable and a model that does not.

The general form for the likelihood ratio test is derived through comparisons of deviances of nested models (see Desideratum 17). The difference in deviances between nested models that differ only in the addition of a single variable approximates a chi-square distribution with one degree of freedom. While the likelihood ratio test is arguably the strongest test of a predictor's statistical contribution to a model, it may be time consuming to fit the appropriate one-variable-added models for each variable in a multivariable LR. The potential intensity of this process is the primary reason why results of the Wald test are often reported and interpreted in manuscripts and reports. However, for studies that do contain continuous predictors or that are based on small samples, results of the Wald tests should be supplemented by the likelihood ratio tests and appropriately included in the Results section of the manuscript.

Given the ready availability of Wald's test statistics, they are sometimes relied on for decisions on trimming non-statistically significant variables from a LR model. Support for these decisions rests

strongly on the adequacy of the data structure as well as on the assumption of appropriate model specification, including the absence or incorporation of salient interactions. That is, all appropriate variables and interactions prior to trimming—from a substantive or theoretical perspective—have been considered. As in the development of all statistical models, variables scientifically critical to the research topic, including relevant interactions, should always be included in the final model, regardless of statistical significance. In situations where the data structure is questionable, decisions on trimming should be supplemented with the likelihood ratio test. Once a final model is decided on, the trimmed model can be compared to the final model using the general likelihood ratio test to assess model fit, described below.

14. Interpretation of Parameter Estimates

The Results section should include interpretation of all effects identified as statistically or substantively significant, and the nature of interactive or polynomial effects should be identified and clarified. As mentioned in the previous Desideratum, parameter estimates in the LR model represent the change in the logit for each one-unit increase in X_j , after controlling for the effects of the other independent variables in the model. Once exponentiated, these regression weights can be interpreted as the odds ratio for the variable ($OR = \exp(\beta_j)$). The further an OR is from 1.0, the stronger the association is judged to be between the independent variable and the outcome (see Desideratum 3).

While strict numerical interpretations of variable effects may be straightforward, there are useful tabular and graphical methods that can aid in understanding the logistic model, particularly in the presence of complex interactions or polynomial effects. Fox and Andersen (2006) describe the use of effect displays for LR models for dichotomous, nominal and ordinal outcomes, and include details on R software code; Jaccard (2001) provides a general process for interpreting interaction effects in LR. Hayes and Matthes (2009) and more recently Hayes (2012, 2013) have developed software to aid in the interpretation of moderation and mediation effects for logistic models and other models. All of the information needed to understand main and/or interaction effects is available from a model's parameter estimates. Given specific values of the independent variables, a logit prediction can be determined and then transformed to an odds and finally to estimated probability of the targeted outcome event occurring. Graphs of estimated logits or predicted probabilities against the levels of a continuous independent variable can conveniently be constructed for a continuum of values of that variable, with multiple lines as needed on the same graph representing different values of a qualitative independent variables. In general, probabilities are more easily interpreted than logits, but either graph could convey the desired information regarding patterns in the data. When an analysis includes more than one continuous predictor, such graphs can become challenging to construct and interpret. However, utilizing these graphical strategies for specific values of a focal predictor can aid greatly in the interpretability of the overall model.

15. Model Assessment

Model assessments should include investigating linearity in the logit for continuous model predictors; corresponding tests of the assumption of proportional or parallel odds for ordinal data; the extent of multicollinearity; the presence of outliers or influential observations (see Desideratum 15); and on locating zero-cells or evidence for separation or complete separation. The goal of these efforts is to develop support for the specification and validity of the model, and these issues should be investigated prior to determination of model fit and interpretation of effects for the final model. A summary of the results of these assessment procedures as well as decisions for dealing with any problems that were identified should be mentioned in the results section of the manuscript.

Linearity in the logit refers to the relation between continuous predictors and the log-odds of the dependent variable. Departure from linearity affects statistical power and tends to underestimate the relation between a predictor and the outcome. In addition to simple graphical depiction, linearity in the logit can be assessed through the Box-Tidwell test or by orthogonal polynomials applied to a categorization of the continuous predictor (Hosmer & Lemeshow, 2000; Menard, 2002). If non-linearity in the logit is evident, adjustments include combining outcome categories, or categorizing continuous independent variables so that separate parameter estimates might be obtained for different values of that predictor.

The proportional or parallel odds (PO) assumption refers to the assumptions regarding the relative stability of variable effects in ordered logistic regression models. For cumulative odds models, the PO assumption implies that the effects of individual predictors remain constant across the underlying binary partitions to the data; similarly for the parallel odds assumption in CR models using the continuation ratio link. These assumptions are tested through a score test, and models allowing non-proportional or non-parallel odds, although more complex, can be developed as needed (Agresti, 2013; Fox, 2008; O'Connell, 2006).

Multicollinearity among predictors tends to inflate the standard errors for the estimated regression coefficients, which in turn affects the validity of statistical tests of these estimates. Multicollinearity also can lead to increased size of the regression coefficients. Multicollinearity is only a property of the predictors; thus, investigations for the presence of multicollinearity proceed as they would with standard linear regression regression. There is currently no single best practice in terms of dealing with multicollinearity in LR, other than its detection followed by review and potential elimination of redundant variables within a multivariable model. Thus, a correlation matrix for the predictors should be included in the results section so that the degree of collinearity can begin to be assessed. A brief discussion of statistical tolerance for the predictors in a multivariable regression model should form part of the Results section. Menard (2002) suggested that as tolerance drops towards .10 (or less), the multicollinearity may be unreasonably large.

The problem of zero-cells occurs when all responses for a particular category of a nominal predictor are exactly the same; this invariance means that one of the two response categories does not occur in the sample data. Depending on the pattern of responses (either all success, or all failure), the estimated logit would be infinitely large or infinitely small, and the standard errors for the estimated logits will be extremely large as well. Options for dealing with zero-cells include collapsing categories of the predictor, eliminating the zero-cell category completely, weighting the data and assigning a particularly small weight to the zero-cell category, or rescaling the nominal variable in some fashion to represent an ordinal variable which can then be included in the model as if it were continuous (with one degree of freedom). These strategies may improve the numerical processes of the model, but will also affect the substantive interpretation of the predictors. An alternative approach uses exact logistic regression methods to determine the parameter estimates based on a conditional likelihood function rather than the standard approach utilizing the ML function (Collett, 2003).

Separation (also referred to as *quasi-separation*) and *complete-separation* are conditions that refer to near perfect or perfect predictions, respectively, of the response variable. As with zero-cells and multicollinearity, standard errors and the coefficients themselves will be extremely large and tend towards infinity. Separation and complete separation tend to occur for smaller sample sizes, and particularly when the number of participants experiencing the event of interest is small (Hosmer & Lemeshow, 2000). The risk of separation increases for increasing numbers of covariates and whenever the number of covariates becomes close to the sample size. Separation is not likely with continuous predictors, but complete separation can occur with any type of data (So, 1995). Although perfect prediction may seem like a great result for any model, separation hinders

understanding of the effects of variables within a model since researchers may wish to determine the effect of a particularly strong variable and under conditions of separation the ML estimate for that regression coefficient no longer exists (the parameter estimate will be infinite). Adjustments for separation or complete separation mirror some of the options for zero-cells such as collapsing or eliminating categories; more advanced strategies are reviewed by Heinze and Schemper (2002), including exact logistic regression. However, they caution that there may be situations in which the requirements of the exact approach may not work well.

Overall, multicollinearity, zero-cells, and separation or complete separation manifest as the same problems within a logistic regression model: inflated standard errors and, often, inflated regression coefficients. Thus, the ability to distinguish among these issues and adequately adjust the analysis for their problematic effects requires generous preliminary assessments of the data. In addition, non-linearity in the logit affects the parameter estimates and thus the interpretation of effects. Thus, it is essential for statistical validity of the final model that any offending variables or combinations of variables be identified, and that all corrective measures are made explicit in the Results section of the manuscript.

16. Residual Diagnostics

Any regression model should be evaluated not only in terms of overall model fit (Desiderata 17 and 18), but also in regard to the distribution of fit (and ill-fit) of the model's predicted values to the values actually observed. These analyses are referred to as residual diagnostics in that they focus on aspects of that which remains "unpredicted" or in error after accounting for the model's predictions. In terms of methodology for residual diagnostics in LR, Pregibon (1981) remains the most influential reference in this field. Additional practical strategies for residual diagnostic evaluation in LR can be found in Hosmer and Lemeshow (2000) and Collett (2003).

Plots of residual statistics and corresponding visual assessment of extreme cases is the evaluation approach that is strongly relied upon during LR, due to the fact that the distributional properties of many of the LR residual statistics are largely unknown (Hosmer & Lemeshow, 2000). Thus, casewise or index-plots of residual statistics can often present a persuasive visual assessment of cases that are poorly fit by a model or that have undue influence in the model's parameter estimates.

In LR, residuals can be defined in terms of a single case or groups of cases that share the same covariate pattern. Software documentation should be reviewed to clarify the method used to calculate case residuals. In addition, standardized or studentized residuals from a LR may not appropriately follow a normal or approximately normal distribution, and researchers should be cautious about interpreting the size of these residual statistics in the same way as one might interpret them from an OLS regression. Similarly, influence statistics such as leverage values or Cook's distance should be evaluated in terms of relative size given values for other cases in the sample. This process of comparing residual statistics across cases can be used to locate extreme or unusual cases based on change statistics, such as the change in deviance or in the regression coefficients when a case (or cases with the same covariate pattern) is removed.

Due to the intensity of residual analyses, specific results or graphs are often not included in a manuscript but the findings from these analyses are discussed more generally. Of import when presenting the results of an LR analysis is the specific mention of having attended to these types of diagnostics and whether or not such evaluations led to manipulation of the data set. Decisions to delete cases on the basis of large residuals or unusual influence on the fitted model should be well-justified. Deletion of extreme cases will necessarily improve the model fit for the sample but unjustified or unexplained deletion on the basis of an extreme residual is unwarranted and tantamount to "stacking the deck" in favor of the given model's fit to that sample's data.

Justification of deletion of extreme cases can include suspected error in data entry or in measurement completion, or conceptual impossibility or improbability of an observed value, but should never rest solely on the size of the residual or influence statistic. A detailed explanation in the Results section should always accompany any manipulation to the raw data. At minimum, results should make specific reference to having completed residual diagnostics, their general results, and any subsequent steps taken.

17. Model Fit and Measures of Association

As described in Desideratum 9, the deviance for a fitted model is defined as $D = -2LL$ and can be thought of as representing “poorness” of fit. A perfect fitted model would have a likelihood of 1 and thus a deviance of 0; values of the deviance farther from 0 represent worse fit. With grouped data (when specific covariate patterns can be grouped together into J finite sets by their frequency of occurrence within the data), the deviance D can be compared to a chi-square test statistic with $J - (p + 1)$ degrees of freedom, where $p + 1$ refers to the number of parameters in the model including the constant. With grouped data or a relatively small number of replicated covariate patterns, the deviance provides a test of goodness of fit. A good fitting model closely reproduces the observed data. In general, if D exceeds the critical chi-square statistic, this suggests that the model does not provide a reasonable representation of the data.

Unfortunately, when the number of unique covariate patterns in the data becomes close to the sample size, which typically occurs when continuous covariates are present, D cannot be assumed to follow a chi-square distribution, not even approximately. However, the difference in deviances between two nested models forms a general likelihood ratio statistic which will follow a chi-square distribution, and model comparisons can be approached from that perspective. Thus, $G = D_{\text{reduced}} - D_{\text{full}}$ is a quantity that represents improvement in fit, where the reduced model contains a subset of variables included in the full model. The degrees of freedom for this general likelihood ratio test is the difference in the number of estimated parameters between the two models. When the reduced model is the null or constant only model, this test is an omnibus test of the fitted model containing p predictors. However, comparing nested models only provides information on whether the model with more parameters yields a statistically significant improvement (reduction) in the deviance relative to the reduced model. Statistical significance could occur even if a better model might still be found. Thus, additional options for describing model fit should be used to supplement the general likelihood ratio test.

The Hosmer–Lemeshow (H-L) test for dichotomous outcomes can be used to approximate a goodness-of-fit test when sparse cells are present in the data (which will nearly always occur when continuous variables are included in the model). The H-L test is based on formation of several groups referred to as “deciles of risk” that represent ordinal groupings of the estimated probabilities from the model. For most samples, ten groups are formed, but there may be fewer groups depending on the similarity of estimated probabilities across different covariate patterns. The frequencies of cases within the deciles are compared to expected frequencies using a Pearson chi-square statistic with degrees of freedom equal to the number of groups (deciles) minus two. If the model fits well, there will be agreement between the observed and expected frequencies, and the null hypothesis of a good fit between the fitted values of the model and the actual data is retained. Extensions to the H-L test are available for ordinal LR models (Pulkstenis & Robinson, 2004).

There have been concerns voiced in the literature regarding the power of the H-L test, but Hosmer and Lemeshow (2000) themselves have advocated that decisions on the adequacy of a model should be supported through a combination of criteria rather than on the results of a single statistical test. In addition to the likelihood ratio test and the H-L test, strategies for considering the quality of a

model also include measures of association similar to R^2 values, and categorical fit measures representing predictive efficiency (Desideratum 18).

There are several logistic regression analogs to the familiar model R^2 from ordinary least squares regression that may be useful for informing about strength of association between the collection of independent variables and the outcome. However, there is some disagreement among researchers regarding which of these pseudo- R^2 measures is best. The likelihood ratio R^2 (R_L^2 ; also called McFadden's R^2) seems to provide the most intuitive measure of improvement in fit and is determined by computing the proportion reduction in deviance obtained from the fitted model (D_m) relative to the null (or empty) model (D_0): $R_L^2 = 1 - (D_m/D_0)$. This statistic is not routinely reported by commercial statistics packages but is easily computed from available output containing deviances for both the null and the full model. Two other measures of association that also are based on model likelihoods and commonly reported in statistical output are the generalized R^2 , also called the Cox and Snell R^2 , and the Nagelkerke R^2 , which in SAS is labeled Max-rescaled R^2 . Menard's (2000) article comparing six different coefficients of determination for logistic regression models is an excellent resource on these pseudo- R^2 statistics for dichotomous outcomes. Hemmert et al. (2016) review eleven pseudo- R^2 statistics including McFadden's and generally support McFadden's approach or related alternatives. However, they also call for researchers to be explicit in identifying the particular pseudo- R^2 presented in applied manuscripts. As there is not agreement among researchers favoring one statistic over another, results of several of these measures of association should be reported and interpreted in the results section of manuscripts.

Finally, information criteria such as Akaike's information criterion (AIC) or Schwarz's Bayesian information criterion (BIC) provide model fit information through different adjustments to the $-2LL$ of a fitted model, based on sample size and the number of predictors. As with the deviance, lower values are more acceptable. These statistics are particularly useful when comparing non-nested models for a set of data. As noted above, however, good model fit in logistic regression is best assessed through a collection of evidence rather than relying on a single criterion.

18. Classification

In addition to measures of association and statistical tests for model fit, quality of a model can also be gauged through classification accuracy. Classification is based on the probabilities estimated from the model, such that if the estimated probability of response for a particular case is greater than .50 (for example), the case is assigned to the "success" outcome; otherwise, it is assigned to the "failure" outcome. Most software packages allow users to choose different cutpoint options for the classification probabilities; two-way tables showing correspondence between actual and predicted outcomes based on various cutpoints can offer additional information about quality of the model. If classification is the goal of the analysis, justification of these decisions must be clearly identified in the Results section. Often, this justification is empirical and based on the relative severity of Type I (classifying a non-event individual as having the event) versus Type II (classifying an event case as a non-event) error. The impact of cutpoint decisions on evaluation of the overall model should be included as part of the Discussion.

Accuracy of classification depends on the match between the classification frequency and the observed frequencies. Note that there are many situations where model fit is considered good but classification may be poor, particularly if there is a low observed percentage for one of the outcome variable groups. Thus, appropriate language should be adopted in presentation and discussion of results. In cases where classification is not reported or is poor, language suggestive of a model's ability to identify or predict group membership should be avoided in preference for language that describes model fit to the sample data.

Percent correct is, by default, reported most often in most major statistical packages, yet is the least effective way in which to describe accuracy of classification because it does not accommodate either base rate or chance classification. Two excellent resources for information on alternative indexes of predictive efficiency are Menard (2000, 2002) and Long (1997). In particular, Menard reviewed and compared an extensive collection of classification measures. Among these are τ for prediction tables, τ_p ; the *adjusted count R²* or R^2_{adjCount} (also called λ_p due to its similarity to the Goodman-Kruskal λ as applied to prediction tables); and, for selection models, ϕ_p . All three of these measures can be interpreted as proportional reduction in error statistics and can be tested for statistical significance. Given a researcher's extensive choice for predictive efficiency measures, it is important that the selected measure or measures be explicitly defined and that the choice matches the nature of the model. For classification models, Menard (2002) suggested that τ_p is the most appropriate measure, and ϕ_p is recommended for selection models. Unfortunately, these statistics are not directly programmed into commercial LR statistics packages and must be calculated from the classification table provided by the software. Whatever choice is made for determination of classification accuracy, sufficient detail and data must be presented in the results section to ensure reasonable and fair interpretation during discussion of the overall effectiveness of the model.

19. Credibility of the Model

The strength and utility of a statistical model rests on more than just global assessments of model fit. To effectively contribute to research and practice, model development must be guided by relevant research questions matched to rigorous methods that are appropriately designed to address those questions. The credibility of all statistical models and the research contributing to their development is, in large part, based on this connection whether for attempting to gather support for a hypothesized causal link or for identifying and describing patterns of associations between variables and an outcome. Thus, reviewers of LR studies should be cautious of methods or results that do not sufficiently address the proposed research questions or hypotheses, or that ignore or brush aside critical issues and factors affecting a LR analysis such as those discussed here. Long (1997, p. 102) argued that measures of fit provide "only partial information that must be assessed within the context of the theory motivating the analysis, past research, and the estimated parameters of the model being considered." Accordingly, it is the researchers' responsibility to situate their work within a well-defined body of literature or preliminary research, to accurately describe their methodology and the measures/variables used, to responsibly justify their statistical decisions when addressing research questions or hypotheses, and to honestly interpret the results and resulting implications in support or refinement of relevant theory. This includes identifying and discussing impacts of limitations to the design, sample, and statistical methodology. Finally, as stated several times throughout this chapter, causal language should only be used when the research design supports such claims.

References

- Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: John Wiley & Sons.
- Agresti, A. (2013). *Categorical data analysis* (3rd ed.). New York: John Wiley & Sons.
- Aldrich, J. H., & Nelson, F. D. (1984). *Linear probability, logit, and probit models*. Newbury Park, CA: Sage.
- Allison, P. D. (1999). *Logistic regression using the SAS system: Theory and application*. Cary, NC: SAS Institute.
- Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 91, 14–28.
- Bring, J. (1994). How to standardize regression coefficients. *The American Statistician*, 48(3), 209–213.
- Clogg, C. C., & Shihadeh, E.S. (1994). *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage.
- Collett, D. (2003). *Modelling binary data* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1–38.
- Dobson, A. J. (2002). *An introduction to generalized linear models* (2nd ed.). Boca Raton, FL: CRC Press.
- Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.
- Finch, W. H., & Schneider, M. K. (2006). Misclassification rates for four methods of group classification: Impact of predictor distribution, covariance inequality, effect size, sample size, and group size ratio. *Educational and Psychological Measurement*, 66, 240–257.
- Fox, J. (2008). *Applied regression analysis and generalized linear models*. Thousand Oaks, CA: Sage.
- Harlow, L. L. (2005). *The essence of multivariate thinking: Basic themes and methods*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Harrell, Jr, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). New York: Springer.
- Hayes, A. F. (2012). *PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling*. White paper. Retrieved from www.afhayes.com/public/process2012.pdf.
- Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis*. New York: Guilford.
- Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior Research Methods*, 41, 924–936.
- Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 22, 2409–2419.
- Hemmert, G. A. J., Schons, L. M., Wieseke, J., & Schimmelpfennig, H. (2016). Log-likelihood-based pseudo-R² in logistic regression: Deriving sample-sensitive benchmarks. *Sociological Methods and Research*, 47(3), 507–531.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley & Sons.
- Huberty, C. (2010). Discriminant analysis. In G.R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 71–78). New York: Routledge.
- Jaccard, J. (2001). *Interaction effects in logistic regression*. Thousand Oaks, CA: Sage.
- King, G., & Zeng, L. (2001a). Explaining rare events in international relations. *International Organization*, 55, 693–715.
- King, G., & Zeng, L. (2001b). Logistic regression in rare events data. *Political Analysis*, 9, 137–163.
- Korn, E. L., & Graubard, B. I. (1999). *Analysis of health surveys*. New York: Wiley.
- Lei, P. W., & Koehly, L. M. (2003). Linear discriminant analysis versus logistic regression: A comparison of classification errors in the two-group case. *Journal of Experimental Education*, 72, 25–49.
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7, 19–40.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. Monographs on statistics and applied probability, no. 37. Boca Raton, FL: Chapman & Hall/CRC.
- McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: John Wiley Sons.
- McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition*. Hoboken, NJ: Wiley.
- Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54, 17–24.
- Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Menard, S. (2004a). Six approaches to calculating standardized logistic regression coefficients. *The American Statistician*, 58(3), 218–223.
- Menard, S. (2004b). Correction. *The American Statistician*, 58, 364.
- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 135, 370–384.
- O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks, CA: Sage.
- Peduzzi, P., Concato, J., Kemper, E., Holzberg, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12), 1373–1379.
- Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, 9, 705–724.
- Pulkstenis, E., & Robinson, T. J. (2004). Goodness-of-fit tests for ordinal response regression models. *Statistics in Medicine*, 23, 999–1014.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177.
- Scott, A., & Wild, C. (2003). Fitting logistic regression models in case-control studies with complex sampling. In R. L. Chambers & C. J. Skinner (Eds.), *Analysis of survey data* (pp. 109–121). New York: Wiley.
- Snyder, L., & O'Connell, A. A. (2008). Event history analysis for communications research. In M. D. Slater, A. F. Hayes, & L. B. Snyder (Eds.), *Sage sourcebook of advanced data analysis methods for communication research* (pp. 125–158). Thousand Oaks, CA: Sage.
- So, Y. (1995). A tutorial on logistic regression. SAS Conference Proceedings: SAS Users Group International 20, April. Retrieved February 13, 2009 from <http://support.sas.com/rnd/app/papers/logistic.pdf>.
- Stefanski, L. A. (2000). Measurement error models. *Journal of the American Statistical Association*, 95, 1353–1358.
- Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston, MA: Allyn and Bacon.
- Taylor, A. B., West, S. G., & Aiken, L.S. (2006). Loss of power in logistic, ordinal, logistic, and probit regression when an outcome variable is coarsely categorized. *Educational and Psychological Measurement*, 66, 228–239.
- Wright, R. E. (1995). Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics*. (pp. 217–244). Washington, DC: American Psychological Association.

17

Log-Linear Analysis

Ronald C. Serlin and Michael A. Seaman

Log-linear analysis is a technique for both exploratory and confirmatory analysis of variable relations when all the variables of interest are classification (i.e., categorical) variables. Researchers who are familiar with contingency-table analysis will recognize log-linear analysis as a tool for assessing higher-order relations among two or more classification variables that cannot be achieved with traditional two-variable chi-square analysis. The term *log-linear* refers to the use of the logarithms of frequencies to create linear models that parallel the linear modeling in analysis of variance (see Chapter 1, this volume). In analysis of variance the mean of the response variable, measured on a quantitative scale, is viewed as a function of effects of explanatory classification variables. For log-linear analysis, it is actually the log of expected frequencies that is associated with factors in a specified explanatory model. Rather than effects, the explanatory variables are the classification variables of interest, so that the test of the model is one of structural relations, rather than effect. A special case of log-linear analysis, logistic regression (see Chapter 16, this volume), has a more direct parallel to analysis of variance, in that there are established response and explanatory variables, as well as a linear function of effects to explain the response. Log-linear analysis involves testing of hypotheses about the interaction among variables of interest. Results provide an understanding of relations and partial relations among these variables. Computations for log-linear analysis can be performed using most major statistical software packages, such as S-Plus, SAS, and SPSS, although the level of user input required varies among the packages. Default settings in software subroutines often do not address hypotheses of interest, so that software use might require a relatively high level of user sophistication. We recommend texts by Agresti (1990), Christensen (1997), and Fienberg (1980). Specific desiderata for applied studies that include log-linear analysis are presented in Table 17.1 and explained in the following sections.

1. Identification of Categorical Factors

The literature review should identify factors in previous studies that are clearly suited for classification of the units of analysis. Often the researcher has a choice of measurement scale, and arbitrary classification of intervals on a quantitative scale can result in categorical analysis that is less powerful for identifying relations than if the original scale was maintained. When the classification is not natural and obvious, the researcher should justify the use of a categorical variable. Similarly,

Table 17.1 Desiderata for Log-Linear Analysis.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. A literature review and statement of purpose refer to factors that are best represented as categorical variables.	I
2. Categorical variables and the levels of these variables are clearly defined.	I, M
3. Hypotheses are proposed regarding relations (or lack of relations) among the categorical variables.	I, M
4. The methods of sampling and data collection are described.	M
5. Log-linear models are developed to correspond to hypotheses of interest.	M
6. Tables are constructed to display frequencies and proportions of cross-classified variables.	R
7. Expected values, odds, and odds ratios are computed.	R
8. Hypotheses are tested by comparing model fit statistics.	R
9. Partial tests of association are conducted to avoid Simpson's paradox.	R
10. The name and version of the utilized software package is reported, along with subroutine choices and justification.	R
11. Planned and post-hoc contrasts are tested using log-odds ratios, where relevant.	R
12. Methods for dealing with cells with sparse or missing data are explicated.	R
13. Effect size measures are reported to assess substantive importance of effects.	R, D
14. Model performance is discussed in the context of theoretical understanding of the factors of interest.	D
15. Results of hypothesis tests are discussed, including unexpected outcomes.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

log-linear analysis does not take advantage of the full information when the classification can be ordered, such as with Likert scales, but is best suited for unordered categorical data. For log-linear analysis to be appropriate, all variables of interest should be unordered classification variables.

To examine relations among the identified factors there must be cross-classification. That is, the researcher should focus on multiple factorially crossed factors that can be identified with a single sample or one or more factors identified for multiple samples (see Desideratum 4). The utility of log-linear analysis is most obvious when there are more than two of these factors. Prior to the development of log-linear methods, the common technique used for studying relations among two categorical variables (contingency table analysis with chi-square statistics) was often used for larger numbers of variables by studying multiple bivariate relations. Studies referenced in the literature review might have used this technique. The statement of purpose in the manuscript should identify a research question about the relations among two or more of the identified factors. The question(s) can be about full or partial associations (see Desiderata 8 and 9). Exploratory studies are typically conducted for the purpose of identifying the structural nature of the relations, but confirmatory studies should clearly identify the nature of the relations.

2. Defining Categorical Variables

Once the factors of interest have been specified, the specific operational variables must be identified. These are unordered categorical variables (see Desideratum 1), so that the operational definitions consist of the classes or categories that will be identified for each unit in the sample. It is important that every unit fits into one and only one category for a single variable. That is, the categories are mutually exclusive and exhaustive. All units must also be amenable to cross-classification on the

other factors to accommodate a complete study of the relations and partial relations. If this is not possible, full study is confined to those units that can be completely cross-classified, and other units can contribute to the study of a more restricted set of questions. If there are too many categories relative to the number of units in the sample, this can result in sparsely populated cells in the cross-classification (see Desiderata 6 and 12). The number of units classified in each category or cell (i.e., cross-classification) can be increased by combining categories, although this will decrease the information that can be deduced from the analysis. For example, collapsing low-frequency categories together into an *other* or *miscellaneous* category will increase the number of units within the category, but some of the information available with the original data is hidden. As with other types of analysis, larger numbers of units at each level of the variable will lead to a more powerful study of those levels, so the researcher must strike a balance between the specificity of the categories and the power available for studying these categories.

3. Proposing Hypotheses Regarding Relations

The manuscript should include hypotheses about the relations among the variables of interest. These hypotheses are of several types, some of which are usually trivial: (1) a hypothesis of equal (or unequal) proportions across all cross-classifications of the factors; (2) a hypothesis of equal (or unequal) proportions across all categories within a single factor (which can be repeated for some or all of the variables); (3) a hypothesis about independence (or association) among two of the factors; (4) a hypothesis about higher-order associations (i.e., more than two factors); and (5) a hypothesis about partial associations (i.e., an association among two or more factors within some or all levels of a separate factor or factors; see Desideratum 9). The most common of these is the hypothesis of association between two factors. Traditional analysis involved chi-square tests of two-factor data, but even when multiple two-factor relations are of primary interest, log-linear analysis provides error control that is typically ignored when multiple two-factor tests are conducted. In exploratory analysis, the hypotheses might not be explicit, but rather the focus is on ascertaining whether relations exist among the factors. For this reason, some common software packages have default settings to conduct all tests or conduct stepwise testing based on hypotheses with statistically significant results. Stepwise approaches are data-driven and thus unreliable and sample-size dependent. If these methods are utilized, cross-validation and replication should be encouraged, if not required. Explicit hypotheses are preferred so that model testing can focus on the questions of interest (see Desideratum 5). Hypotheses are frequently stated in terms of research hypotheses, rather than statistical hypotheses.

4. Describing Sampling and Data Collection

There are three common sampling models that result in data collection best suited for log-linear modeling: (1) multinomial sampling, (2) product-multinomial sampling, and (3) Poisson sampling. Multinomial sampling is the process of selecting a sample of individuals from a single population and then cross-classifying every member of the sample on two or more factors. Product-multinomial sampling is the process of selecting a fixed number of units from each of two or more populations that differ in terms of classification on a single factor. Each unit within each of the multiple samples is then classified on the basis of one or more additional factors. The result of data tabulation from these two sampling methods looks identical: cross-tabulated frequency data for multiple factors, with each unit represented in one and only one cross-classification. The Poisson sampling model, though less common in social research, also results in data that can be analyzed using log-linear analysis. In this model, the possible

cross-classifications, but not the number of observations, are known in advance. Observations during a fixed time period yield cell frequencies. These three sampling models do not form an exhaustive list, but they do encompass most of sampling types that lead to cross-tabulated data in published research.

A serendipitous property of these sampling models is that all three types asymptotically lead to the same log-linear analysis (Feinberg, 1980). More advanced so-called *exact methods*, which derive confidence sets or critical values from permutations of observations, differ for each sampling method, but these methods are rarely found in the applied literature and will not be discussed here. Asymptotic methods are reasonable when the standard rule of thumb is observed that no more than 20% of the cross-classified cells have expected values less than 5 (see Desideratum 12 for more precise sample size rules).

Researchers often do not discuss the sampling method, presumably because the common sampling models asymptotically lead to the same statistical results. This is problematic. It is important to explicate the method, because the interpretation of the findings is directly linked to the sampling method. For example, in one model (product-multinomial sampling) the findings relate to the homogeneity (or lack thereof) of multiple populations on one or more factors. By contrast, when the sampling is multinomial the findings refer to the relations of factors for a single population.

An advantage of these types of data collection is that there is often little or no ambiguity regarding the classification of the unit on the factors of interest. Thus, validity of the “measurement” is not an issue. If classification is ambiguous, construct validity is an issue that must be addressed. The researcher must clearly define classifications that are not widely known and accepted. If units are classified by judges, interrater reliability must be established (see Chapter 10, this volume). If classifications are created using intervals on a scale of otherwise ordered categories, the researcher must defend the choice of unordered classification, because typically this choice results in less specific results and/or a loss of power.

5. Log-Linear Models and Hypotheses of Interest

As with all research, the proposed hypotheses’ truth or falseness is assessed by examining how well the collected data accord with them. The relation (or lack of relation) among measures of theoretical constructs that is specified in a hypothesis posits that a nonzero (or zero) value, respectively, of a log-odds or log-odds ratio exists in the population sampled. Analogous to analysis of variance, log-linear analysis is used to evaluate whether the data seem unlikely to have been sampled from a population in which the null hypothesized relations exist; if the data seem unlikely, given the hypotheses, then the null hypotheses are declared false. The hypothetical relations among the variables that are examined in the study must be clearly delineated in the manuscript.

The determination of the likelihood of the data, given the relevant hypotheses, is achieved through the use of a linear model. As in analysis of variance, the model includes an intercept (or grand mean) and main effect and interaction terms that are consistent with the specified hypotheses. For instance, in analysis of variance, one can test whether a population’s mean on a continuous dependent variable is equal to a theoretically specified value. In log-linear analysis, because the variables are categorical, the analogous test would examine whether the log-odds, the logarithm of the ratio of the probabilities of the population falling into one category or another, is equal to a theoretically specified value.

In both instances, the data are speculated to arise as if generated at random according to an underlying linear model that includes terms whose magnitudes are theoretically specified by the hypotheses of interest. This model must be clearly delineated in the research report. In the example of analysis of variance, a null hypothesis regarding a population mean is specified as

$$H_0: \mu_t = \mu_0,$$

where μ_t denotes the true expected value of the population from which the sample was drawn, and μ_0 denotes the theoretically specified expected value of the population. Equivalently, in log-linear analysis the null hypothesis concerns log-odds, the ratio of probabilities of members of the population falling into one or another category of the dependent variable, leading to the null hypothesis

$$H_0: \ln O_{t_{i_1 i_2}} = \ln O_{0_{i_1 i_2}},$$

where $\ln O_{t_{i_1 i_2}}$ denotes the ratio of the true probability of members of the population falling into category i_1 or i_2 , and $\ln O_{0_{i_1 i_2}}$ denotes the theoretically specified ratio of the probability of members of the population falling into category i_1 or i_2 .

With a specified sample size, N , the null hypothesis can equivalently be written in terms of *expected frequencies*, F_i , where $F_i = Np_i$, where p_i denotes the probability of a member of the population falling into category i . In particular, log-linear analysis deals with models that specify logarithms of expected frequencies in terms of values of parameters that are determined by the hypothesized relations.

Most generally, if the hypotheses specify relations among variables (or factors) A, B, C, . . . , then the log-linear model would be written

$$\ln p_{ABC\dots} = \mu + \alpha_A + \alpha_B + \alpha_C + \dots + \gamma_{AB} + \gamma_{AC} + \gamma_{BC} + \dots + \gamma_{ABC} + \dots$$

where in direct analogy with analysis of variance, μ denotes a grand mean, α denotes a main effect, and γ denotes an interaction. As main effects, the α terms represent differences in logarithms of probabilities, equivalent to the logarithms of ratios of probabilities, or log-odds. As interactions, the γ terms represent differences of differences of logarithms of probabilities, equivalent to the differences of logarithms of ratios of odds, or differences of log-odds ratios. The hypothetical relations among the variables must be clearly specified in terms of parameters in the log-linear model, whereby those log-odds and log-odds ratios that are nonzero according to the hypothesized relations must correspondingly have nonzero effects in the model, and those log-odds and log-odds ratios that are zero according to the hypothesized relations must have the corresponding effects set to zero, thereby not having those terms included in the model.

6. Cross-Classification Tables

A *cross-classification table*, also known as a *crosstabs* or *contingency table*, is a display of the frequencies or proportions for all of the possible cross-classifications. The number of cross-classifications is simply the product of the numbers of categories identified for each factor. Table 17.2 contains a sample table for three factors, with two categories for two of the factors and four categories for the third factor.

It is more useful, alternatively or additionally, to display proportions or percentages, where the proportion is defined as f / N_h . Here f is the frequency for a category or a cross-classification cell and N_h is the appropriate sample size for the hypothesis of interest. It is unfortunately common for researchers to ignore the hypothesis (and therefore the research question) of interest when creating a cross-classification table. Proportions should be clearly defined and match interest. For example, using the total sample size for N_h is misleading when the question is one of homogeneity of multiple populations on a response variable. In this case, N_h should be the individual sample sizes for the samples drawn from each of the populations. Researchers should be clear about defining

Table 17.2 A $2 \times 2 \times 4$ Cross-Classification Table.

	Factor A, Category 1		Factor A, Category 2	
Factor C	Factor B, Category 1	Factor B, Category 2	Factor B, Category 1	Factor B, Category 2
Category 1	f_{111}	f_{121}	f_{211}	f_{221}
Category 2	f_{112}	f_{122}	f_{212}	f_{222}
Category 3	f_{113}	f_{123}	f_{213}	f_{223}
Category 4	f_{114}	f_{124}	f_{214}	f_{224}

the proportions in the table, rather than requiring the reader to infer this by looking for the set of proportions that sum to one. The researcher should also make use of the descriptive information provided by the cross-classification table prior to focusing on the model-testing results. A well-constructed cross-classification table clearly leads to preliminary results, such as homogeneity (or heterogeneity) of samples, independence (or association) of factors, and partial independence (or association) of factors.

7. Expected Values, Odds, and Odds Ratios

Tests of log-linear models are based on the values one might expect to appear in cross-classification tables if the model is correct. Various terms of the model suggest contributions to the expectation of frequencies within categories or cross-classifications of categories. Essentially, a log-linear model should be retained or rejected depending on how close a match exists between the expectations provided by the model and the actual frequencies observed in the data. The calculation of an expectation is based on the marginal frequencies in the data and the hypothesis of interest, as reflected by terms in the model. For example, the common hypothesis of independence of two factors suggests that the expectations for the cross-classification of these factors can assume factor independence. Thus, $E(f_{ij}) = f_i p_i p_j$, where $E(f_{ij})$ is the expectation of the frequency for the i th category of Factor A and the j th category of Factor B given the total sample size (f) and the proportion of the units that are in category i of Factor A (p_i) and category j of Factor B (p_j). Similar calculations are available for both simpler and more complex hypotheses, though the complexity of the calculation for higher-order interactions and multiple interactions might require iterative processes.

Interactions should be descriptively examined using odds and odds ratios. An estimate of the odds of appearing in one category is given by p_1/p_0 , where p_1 is the proportion of units in the category and p_0 is the proportion of units not in the category. With count data, interactions among factors are examined with odds ratios using contingent odds. That is, the odds of a unit appearing in a category for one factor might be contingent on the category the unit is in for a second factor. Ratios of these odds for different categories of one of the factors provide an estimate of the strength of association of the two factors.

Table 17.3 illustrates a simple two-factor design with two categories in each factor. The estimate of odds of being in Category 1 of Factor B for units in Category 1 of Factor A is f_{11}/f_{21} . Similarly, the estimate of odds of being in Category 1 of Factor B for units in Category 2 of Factor A is f_{12}/f_{22} . The odds ratio is then $(f_{11}/f_{21})/(f_{12}/f_{22})$. If the odds of being in Category 1 of Factor B are not contingent on the categories of Factor A, then the odds will be the same and the ratio will be 1. Departures from 1 indicate a relation between Factors A and B, and the distance from 1 indicates the strength of this relation. With more factors and categories, contrasts of odds ratios should be used to study specific interactions related to the research questions.

Table 17.3 Frequencies in a Two-Factor Study.

		<i>Factor A</i>		
		<i>Category 1</i>	<i>Category 2</i>	
Factor B	Category 1	f_{11}	f_{12}	$f_{1\cdot}$
	Category 2	f_{21}	f_{22}	$f_{2\cdot}$
		$f_{\cdot 1}$	$f_{\cdot 2}$	$f_{\cdot \cdot}$

8. Hypothesis Tests and Model Comparison

In order to test hypotheses of interest, test statistics are computed to assess how well the observed frequencies, $f_{ijkl\dots}$, match the expected frequencies, $F_{ijkl\dots}$. The two statistics most commonly used for this purpose are Pearson's chi-square statistic,

$$\chi^2 = \sum_{i,j,k,\dots} \frac{(f_{ijk\dots} - F_{ijk\dots})^2}{F_{ijk\dots}},$$

and the likelihood ratio statistic,

$$G^2 = 2 \sum_{i,j,k,\dots} f_{ijk\dots} \ln \left(\frac{f_{ijk\dots}}{F_{ijk\dots}} \right),$$

both of which approximately follow (with large samples) a chi-square distribution. Research has shown that in most small-sample conditions, the Pearson statistic controls the Type I error rate better than the likelihood statistic. Nevertheless, when testing the effects of specific hypothesized relations, as reflected by tests of model parameters, the additive property of the likelihood statistic suggests that tests based on G^2 should be conducted and values of G^2 and associated p -values should be reported, as is typically done in log-linear analysis.

Both χ^2 and G^2 are known as *goodness of fit* statistics, in that they quantify how well a model seems to fit the data, but in log-linear analysis they might better be referred to as "badness-of-fit" statistics, in that they get larger (and p -values smaller) as the fit between model and data becomes worse. This is especially salient in log-linear analysis, because there are many potential models (e.g., with 3 variables there are 19 models, not counting the model containing only the grand mean), and many of these models yield a statistically significantly bad fit between observations and expectations. If one focused solely on the fit between models and data, one would somehow have to determine that one significantly bad model is better than another significantly bad model. Furthermore, one would be searching for a model whose G^2 indicates a lack of significantly bad fit, which would lead a researcher to attempt to draw a conclusion on the basis of the lack of significance of a null hypothesis test, well known to be logically invalid.

Finally, it is only by focusing on the comparison in pairwise fashion of a much smaller set of models that one can draw conclusions about the original hypotheses of interest. For example, consider an investigation that involves only two variables, Factors A and B, and assume that one's hypotheses of interest involve both of the main effects and the interaction. Then there are five models possible:¹

$$\ln F_{ij} = \mu$$

$$\ln F_{ij} = \mu + \alpha_A$$

$$\ln F_{ij} = \mu + \alpha_B$$

$$\begin{aligned}\ln F_{ij} &= \mu + \alpha_A + \alpha_B \\ \ln F_{ij} &= \mu + \alpha_A + \alpha_B + \gamma_{AB}\end{aligned}$$

If the fit of the second model is statistically significantly better than the first, then one would conclude that the main effect of Factor A is nonzero in the population, because the fit of a model in which this term is nonzero has fit the data better than one in which the term is zero. It is here that the additive property of the G^2 statistic becomes essential, because the difference in the two models' G^2 statistics is itself distributed as chi-square. By comparing the reduction in G^2 to a critical value obtained from the chi-square distribution, one can test the null hypothesis that the A main effect is zero, that is, $G_A^2 = G_\mu^2 - G_{\mu+\alpha_A}^2$, where G_A^2 is a sample statistic testing the null hypothesis $H_0: \alpha_A = 0$, and G_μ^2 and $G_{\mu+\alpha_A}^2$ are likelihood ratio statistics assessing the goodness of fit of log-linear models including only the grand mean, μ , or including both the grand mean and the main effect of Factor A, α_A , respectively.

One can test the main effect of B and the interaction of Factors A and B in a similar fashion. The last of these models, which includes all of the parameters, is known as the *saturated* model. In a saturated model the expected frequencies are equal to the observed frequencies, and so the G^2 statistic is equal to zero. The test of the highest-order interaction is identical to the test of the "goodness" (badness) of fit of the model in which all lower-order terms are included. The G^2 statistics used for testing the hypotheses of interests and their associated p -values should be reported in the manuscript.

9. Marginal and Partial Tests and Simpson's Paradox

Referring again to the two-factor example and the five associated models of interest (see Desideratum 8), notice that the main effect of Factor A can be tested by comparing the fit of the second and first models, as described, but also that the main effect of Factor A could be tested by comparing the goodness of fit of the fourth and third models, the former containing the main effect of A and the latter not. In the case of testing main effects, both approaches yield identical results, regardless of how many factors are included in the design. The same cannot be said when testing interactions in designs involving three or more factors. In such cases, one must distinguish between two approaches.

For one of these approaches, a *marginal* test of the interaction under examination, the researcher compares the G^2 statistic for a model in which only the interaction in question and the requisite lower-order terms are included to the G^2 statistic for a model that includes the same lower-order terms but not the interaction under examination. This test is conducted as if all factors not included in the interaction do not exist in the design, and the test is performed as if the data table had been collapsed across all other factors. For example, consider testing the hypothesis that the AB interaction is zero in a design that includes three factors, A, B, and C. The marginal test statistic would be calculated as $G_{AB}^2 = G_{\mu+\alpha_A+\alpha_B}^2 - G_{\mu+\alpha_A+\alpha_B+\gamma_{AB}}^2$, comparing the goodness of fit of two models that do not include Factor C.

The other approach, a *partial* test of the interaction, includes all of the factors in the design. The partial test is conducted by comparing the G^2 of the model in which all interactions of the same order as the one of interest are included, to the G^2 of the model in which all interactions of the same order as the one of interest, but not the interaction of interest itself, are included, or $G_{AB,C}^2 = G_{\mu+\alpha_A+\alpha_B+\alpha_C+\gamma_{AC}+\gamma_{BC}}^2 - G_{\mu+\alpha_A+\alpha_B+\alpha_C+\gamma_{AB}+\gamma_{AC}+\gamma_{BC}}^2$, where $G_{AB,C}^2$ represents the likelihood ratio test of the AB interaction from which Factor C has been partialled, $\gamma_{AB,C}$.

Table 17.4 Example Illustrating Simpson's Paradox.

		B_1	B_2
C_1	A_1	30	180
	A_2	70	220
C_2		B_1	B_2
	A_1	350	90
	A_2	150	10

Partial tests should always be conducted and reported, because marginal tests may fall prey to what is known as Simpson's paradox, resulting in erroneous conclusions due to the confounding influence of other variables that the marginal tests exclude from the analysis. As an example, consider again the design that includes Factors A, B, and C, and assume that each factor has two levels and that the data are as shown in Table 17.4. In level C_1 , the odds of falling in level A_1 versus level A_2 equal 0.4286 for level B_1 and 0.8182 for level B_2 , and similarly in level C_2 , the corresponding odds equal 2.3333 for level B_1 and 9 for level B_2 ; that is, in both levels of C, the odds are greater in level B_2 than in level B_1 , and they are statistically significantly so. Yet, if one collapsed the table across levels of C, one would find that the corresponding odds in level B_1 , 1.7273, are greater than the odds in level B_2 , 1.1739, and statistically significantly so! Thus, one would draw a conclusion from the collapsed, marginal table that is opposite those found in the separate levels of variable C; this is an instance of Simpson's paradox.

The explanation for such an occurrence is that variable C is associated with both variables A and B, and this so-called third-variable influence can result in a spurious result if it is not accounted for. When the odds are compared in the separate levels of variable C, this variable's influence is controlled, or *partialed*, and the relationship between A and B thus cannot be due to variations in C. Because it is desirable that our conclusions be unconfounded, the tests reported in log-linear analysis should examine partialled effects.

10. Computer Programs

Researchers should specify which statistical computer program was used to perform the analyses and which default and custom options were selected, in order to indicate how estimates and test statistics were obtained. Most statistical packages perform log-linear analysis, with the notable exception of Minitab, and some programs offer several routines that do so. For instance, the SPSS routines GENLOG, HILOGLINEAR, and LOGLINEAR can all be used in this endeavor (although LOGLINEAR can only be accessed through a syntax window), and in SAS both PROC CATMOD and PROC GENMOD can yield log-linear analyses.

If a researcher is interested in testing partialled effects, then the procedure to use in SPSS is HILOGLINEAR, selecting “Model Selection” from the Analyze>Loglinear tabs and choosing the “Association Table” option with a default saturated model. SPSS adds a constant, denoted delta and equal to 0.5, to all cells when performing this analysis, and the researcher is encouraged to set delta equal to a very small number, say 0.00001. These options should be delineated in one’s manuscript. To test partialled effects in SAS, PROC GENMOD should be used. A saturated model should be specified with a Poisson distribution, a LOG link, and a TYPE3 analysis should be requested. Again, these specifications should be indicated in the manuscript.

11. Planned and Post Hoc Contrasts

In analysis of variance, a statistically significant F statistic whose numerator degrees of freedom exceed unity leads one to infer only that the complex null hypothesis under examination is false—that the means, for example, differ but not in precisely what way. Similarly, a statistically significant G^2 test of a main effect or interaction with degrees of freedom of two or more does not indicate the manner in which odds ratios differ from unity or from one another. In both kinds of analyses, planned or post hoc comparisons are examined to provide more detailed information regarding the model parameters and, consequently, about the research questions that the analysis is intended to address.

Analogous to analysis of variance interaction contrasts, written in terms of differences of mean differences, contrasts in log-linear analysis constructed to examine interactions among design factors are written in terms of differences in differences of logarithms of population proportions and are tested in terms of corresponding sample statistics. As an example, say a design contains two factors, Factor A having two levels and Factor B having three levels. Then the complex null hypothesis would be written as $H_0: \gamma_{AB} = 0$. This hypothesis is equivalent to $H_0: \gamma_{(1,2)(1,2)} = \gamma_{(1,2)(1,3)} = \gamma_{(1,2)(2,3)} = 0$, where the first pair of subscript values indicates that both levels of Factor A are involved in the interaction contrast, and the second pair of values indicate that particular levels of Factor B are involved. For example, we would write the first of these interactions as

$$\gamma_{(1,2)(1,2)} = \ln p_{A_1B_1} - \ln p_{A_1B_2} - (\ln p_{A_2B_1} - \ln p_{A_2B_2}).$$

It has been shown (Gart & Zweifel, 1967) that this interaction is best estimated in terms of sample cell frequencies as

$$\hat{\gamma}_{(1,2)(1,2)} = \ln(f_{A_1B_1} + .5) - \ln(f_{A_1B_2} + .5) - [\ln(f_{A_2B_1} + .5) - \ln(f_{A_2B_2} + .5)],$$

and that the variance of this contrast is best estimated as

$$\sigma_{\hat{\gamma}_{(1,2)(1,2)}}^2 = \frac{1}{f_{A_1B_1} + .5} + \frac{1}{f_{A_1B_2} + .5} + \frac{1}{f_{A_2B_1} + .5} + \frac{1}{f_{A_2B_2} + .5}.$$

According to Goodman (1964), in its most general form the omnibus null hypothesis is equivalent to writing that all linear combinations of tetrad interaction contrasts are equal to zero.

Most commonly, interaction contrasts are most interpretable when they involve a fourfold table. Consider again the data presented in Table 17.4. The contrast that would be used to test the partialled AB interaction would be written as

$$\begin{aligned}\hat{\gamma}_{ABC} &= \frac{1}{2}[\hat{\gamma}_{(1,2)(1,2)(1)} + \hat{\gamma}_{(1,2)(1,2)(2)}] \\ &= \frac{1}{2}[\ln(30.5) - \ln(180.5) - \ln(70.5) + \ln(220.5) + \ln(350.5) - \ln(90.5) - \ln(150.5) + \ln(10.5)] \\ &= \frac{1}{2}[(-0.6377) + (-1.3086)] \\ &= -0.9731\end{aligned}$$

The variance of this contrast would be calculated as

$\sigma_{\hat{\gamma}_{ABC}}^2 = \frac{1}{4}[\sigma_{\hat{\gamma}_{(1,2)(1,2)(1)}}^2 + \sigma_{\hat{\gamma}_{(1,2)(1,2)(2)}}^2] = 0.0432$, and the test of this interaction, which is asymptotically normally distributed, would equal $z_{\hat{\gamma}_{ABC}} = \frac{-0.9731}{\sqrt{0.0432}} = -4.6816$. Standard multiple comparison

procedures can be applied to tests of log-linear contrasts. Goodman (1964) showed that Scheffé's method is applicable, with the associated contrast tests having appreciably lower power than tests of the same contrasts using other multiple contrast procedures such as the planned contrast methods due to Dunn, Holm, or Shaffer. Dunn's procedure provides less power than the Holm sequentially rejective method, which in turn has lower power than the improved sequentially rejective method due to Shaffer. Except in those cases for which all but one of the factors possess two levels, the Shaffer method is much more difficult to apply than the Holm procedure, and so in general one would do best to rely on the Holm method to control Type I error rate. Regardless of choice of multiple comparison procedure, interpretable log-odds ratios should be tested and reported.

12. Sparseness

Traditionally, and correctly, concern has been expressed about the adequacy of the chi-square distribution to approximate tail probabilities of statistics when sample sizes are not large. Over the last 70 years, a number of suggestions have been offered regarding the minimum expected cell frequencies required for the approximation to be reasonably good (with suggested minima ranging between 1 and 20). More recently, however, researchers have found that both the magnitude of the expected frequencies and the ratio of total sample size to the number of cells involved in the interaction under study seem to relate to the adequacy of the approximation to the distribution of either the Pearson chi-square or the likelihood ratio statistics in sparse tables (i.e., those in which there are a number of cells with small frequencies). For instance, Agresti and Yang (1987) suggested that the chi-square approximation to the distribution of the likelihood ratio statistic performs poorly when testing the fit of a log-linear model when tables have cells with small expected frequencies, and that the chi-square approximation to the distribution of the Pearson statistic is adequate when $N \geq 10\sqrt{K}$, where N is the total sample size and K is the number of cells in the table. When testing hypotheses regarding partialled effects via subtraction of likelihood ratio statistics, however, these authors found that the chi-square distribution worked well in an $r \times c \times k$ table in approximating the distribution of the likelihood statistic for partialled tests when $N > 5[\max(rc, rk, ck)]$. On the other hand, the chi-square approximation to the distribution of the difference in Pearson fit statistics was inadequate throughout. The results also showed that adding a constant to all cell frequencies, typically 0.5, made both the Pearson and likelihood ratio statistics overly conservative for larger tables in testing both the fit of the log-linear model and the statistical significance of partialled effects (recall, however, that adding 0.5 to cell frequencies in the estimation of log-odds ratios and their variances yields unbiased estimators). Authors should comment on the likely adequacy of the chi-square approximation when testing the fit of the log-linear model, and they should note whether a constant has been added to cell frequencies in calculating the test statistics.

13. Power Analysis and Measures of Effect Size

In addition to reporting the results of statistical tests of parameters of interest, an assessment of the substantive importance of the design factors should be provided. There are a number of such measures available, but it might be most useful to provide measures that are directly analogous to the more familiar measures of effect size used to describe results of an analysis of variance or a multiple regression analysis. When assessing the importance of an effect in analysis of variance, the measure summarizing the size of mean differences among two or more groups is η^2 , the ratio of the between-group and total variances. When assessing the strength of the relation between two variables, one often uses Pearson's correlation coefficient, and the natural extension to the relation between a set of predictors and a dependent variable yields the squared multiple correlation coefficient, R^2 . These

measures have analogs in log-linear analysis. Indeed, the measures in log-linear analysis have the same use in calculating prospective power of a test as do their counterparts in analysis of variance, so that the results of the power calculation can be reported in the manuscript.

The underlying conceptualization of the measures of effect size can be presented in terms of the Pearson chi-square statistic and the measure of the effect size for a contingency table, Cramer's V , whereby $\chi^2 = NV^2$. The measure analogous to η^2 can be defined similarly in terms of the likelihood ratio statistic, namely, $G^2 = N \left[2 \sum \hat{p}_{1k} \ln(\hat{p}_{1k} / p_{0k}) \right] = N \hat{\eta}^2$, where \hat{p}_{1k} is the observed value of a cell proportion and p_{0k} is the value of the cell proportion specified under the truth of the null hypothesis. Clearly, the easiest way to calculate this effect size measure is to divide the G^2 test of a factor by N .

In similar fashion, a measure analogous to the Pearson contingency coefficient can be obtained from the same relationship for a four-fold table. More particularly, the specific relationship between a one-degree of freedom χ^2 variate and the square of a standard normal variate is $\chi_1^2 = z^2 = N\phi^2$, where ϕ is the Pearson contingency coefficient, and $z = \gamma / \sigma_\gamma$. From this, an analog to ϕ is given by

$$\phi_r = \frac{\ln[(p_{11}p_{22})/(p_{21}p_{12})]}{\sqrt{\frac{1}{p_{11}} + \frac{1}{p_{21}} + \frac{1}{p_{12}} + \frac{1}{p_{22}}}}.$$

These measures of effect size relate in a natural way to power calculations, allowing a researcher to calculate the sample size required to detect an effect of a particular magnitude with sufficient power. In the case of the test of a main effect or interaction via G^2 , one would need to specify the true probabilities in each cell of the design, p_{1k} , and the probabilities expected in each cell under the null hypothesis, p_{0k} , resulting in a noncentrality parameter λ , where

$$\lambda = N \left[2 \sum p_{1k} \ln(p_{1k} / p_{0k}) \right].$$

Similarly, when performing a power analysis for a test of a log-odds ratio in a fourfold table, one would calculate λ as

$$\lambda = N \left[\frac{\ln^2[(p_{11}p_{22})/(p_{21}p_{12})]}{\frac{1}{p_{11}} + \frac{1}{p_{22}} + \frac{1}{p_{12}} + \frac{1}{p_{21}}} \right].$$

These parameters would be used to perform the power analysis using a computer program such as G*power or NCSSCALC.

14. Model Performance Linked to Theory

The factors of interest are originally conceived as theoretical constructs that represent broad ideas and definitions about what underlies observable outcomes. These constructs are not themselves observable, but are inventions that define the researcher's understanding. The choice of specific variables moves the study from theory to practice and the definitions of mutually exclusive and exhaustive categories for each of the variables gives full operational form to the study. Log-linear models posit relations among the variables. When models do not fit well with the observations it is because terms are absent from the model that would have changed the expectations for cell frequencies. The relative importance of missing terms can be examined by determining how the addition of a term changes the fit of a model. These salient terms indicate operational variable relations that translate into relations among theoretical constructs. Researchers should describe both model

performance and what this suggests about the factors of interest. This link between the outcomes of the study and theoretical understanding can be strengthened through reference to similar outcomes in other studies, but is restricted in a single study by both sampling and ecological considerations, primarily because of the inherent specificity that must be associated with the operational form of a single study.

15. Discussion of Hypotheses, Outcomes, and Expectations

The discussion about study findings will differ for exploratory and confirmatory studies. In exploratory studies, most or all models will be tested and the emphasis will be on terms that most contribute to fit of the model. A hypothesis of model fit will be rejected when these terms are absent from the model, thus suggesting a specific interaction among some or all of the variables. The exploratory nature of the study compels the researcher to discuss both statistically significant and non-significant contributions to the model. Statistical significance is a function of sample size and the reliability of the categorizations, so it is important for the researcher to also consider effect size (i.e., strength of associations) when discussing the results (see Desideratum 13). Statistically non-significant terms in model fit do not confirm the lack of relations, nor do statistically significant terms establish importance without the assessment of the size of relations. In confirmatory studies the focus is on specific models that highlight relations of interest. The researcher should test those models that highlight relations of interest and should discuss test outcomes regardless of whether results are anticipated or anomalous. In the case of anomalies, the researcher should provide further discussion and suggestions for future studies that would address the theoretical problems posed by the findings. For all relations of a higher order than two factors, the researcher should be careful to distinguish among non-contingent and contingent relations (see Desideratum 9).

Note

- 1 It might seem that there are other models possible, such as $\ln F_{ij} = \mu + \alpha_A + \gamma_{AB}$, but log-linear models are *hierarchical* in nature, so that an interaction term such as γ_{AB} cannot enter a model unless all lower level terms, here both main effects, are also entered, because an interaction cannot be defined without reference to the associated main effects.

References

- Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.
- Agresti, A., & Yang, M.-C. (1987). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics & Data Analysis*, 5, 9–21.
- Christensen, R. (1997). *Log-linear models and logistic regression* (2nd ed.). New York: Springer.
- Fienberg, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge, MA: MIT Press.
- Gart, J. J., & Zweifel, J. R. (1967). On the bias of various estimators of the logit and its variance with applications to quantal bioassay. *Biometrika*, 54, 181–187.
- Goodman, L. (1964). Simultaneous confidence limits for cross-product ratios in contingency tables. *Journal of the Royal Statistical Society, Series B*, 31, 486–498.

18

Mediation and Moderation

Paul E. Jose

Statistical mediation and moderation are two frequently used analytical methods for examining the associations among at least three variables. The conceptual underpinnings, computational methods, and interpretational approaches of mediation will be presented in the first part of this chapter, and following this, the same three themes will be examined for moderation. The chapter will end with a brief description of hybrid models that incorporate both mediation and moderation, termed *conditional process analysis* by Hayes (2013).

1. Mediation

1.1. The Mediation Hypothesis

The authors should conceptualize their mediation model as a temporal prediction (or a causal argument). Although key readings in the field of mediation (Baron & Kenny, 1986; Hayes, 2009; Kenny, 2008, 2014; MacKinnon, 2008) make the point that a mediational model is an explicit causal argument (i.e., X causes Z, which, in turn, subsequently causes Y), researchers sometimes inappropriately couch their hypothesis in inappropriate language.

For single-occasion mediation, researchers may be allowed to use the word “predicts,” as in “the IV was hypothesized to predict the mediating variable, which, in turn, was expected to predict the DV.” Stronger language, such as “causes,” “impacts on,” or “influences” should be reserved for longitudinal and experimental datasets. The field generally allows the use of causal language in experimental designs, whereas in subject variable designs (i.e., no random assignment, no manipulation) arguments of causality are weaker and hence more cautious language may be more appropriate.

A proper mediation hypothesis links all three variables together in a causal or temporal chain, for example, “the cognitive behavioral therapy intervention was predicted to lead to a decrease in irrational thoughts at the conclusion of training, and this diminished level of irrational thinking was predicted to lead to a reduction in depressive thoughts and cognitions six months later.” Many hypotheses refer to “explanation of variance,” as in “the mediator of gratitude explained significant variance in the relation between mindfulness practice (IV) and happiness (DV).” This approach is technically correct but does a poor job of describing the direction of influences. The essence of a mediation hypothesis is to pose a possible mechanism by which the independent variable (IV) affects the dependent variable (DV).

Table 18.1 Desiderata for Mediation and Moderation.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. Mediation	
1.1 The mediation hypothesis: construction of the mediation model is based on an understanding that it will examine temporal prediction, and possibly causal influences, among the variables included in the model.	I
1.2 Types of variables are appropriate for placement within the mediation model.	M
1.3 The dataset is appropriate for the proposed mediation model.	M
1.4 The method of data analysis to derive an estimate of the indirect effect is sensitive, accurate, and unbiased.	M
1.5 The statistical outputs of the mediation analysis are appropriate and reported clearly.	R
1.6 Errors in computation (e.g., multicollinearity) are avoided.	R
1.7 Effect size estimates are obtained and interpreted properly.	R
1.8 Interpretation of the mediation result is appropriate for the types of variables and data used.	D
2. Moderation	
2.1 Examination of a moderation effect is based on an understanding of what a statistical interaction tells one about groups of individuals in one's dataset.	I
2.2 Use of moderation is based on a clear awareness of how it differs from the approach of mediation.	I
2.3 The types of variables are appropriate for their respective roles in the model.	I
2.4 Preparation of the interaction term is performed correctly.	M
2.5 The moderation analysis is performed optimally and without bias (i.e., multicollinearity)	R
2.6 Graphing the moderation result is performed in such a fashion as to correctly and clearly depict the statistical finding.	R
2.7 Simple slope analyses are performed correctly in order to elucidate the statistical finding.	R
2.8 Interpretation of the moderation result is: based on the figure, clear and accessible, and relevant to the proposed hypothesis.	D
3. Process analysis	
3.1 The proposed process analysis model appropriately combines both mediation and moderation.	I
3.2 The types of variables are appropriate for the model proposed.	I
3.3 The model is analyzed with an appropriate statistical program, given the nature of the variables and the dataset.	M
3.4 Post-hoc probing of moderation is performed correctly.	R
3.5 Explanation of the model properly incorporates both mediation and moderation terminology.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

1.2. Types of Variables

Authors are not always aware of constraints on types of variables placed within the mediation model. In most mediation models the IV, the mediator (MedV), and the DV will be continuous. It is permissible for the IV to be dichotomous categorical (e.g., gender). If either (or both) of the MedV or the DV is/are categorical, then their respective regressions will need to be logistic in computation. These design variations, sometimes called *logistic mediation*, are legitimate models that can be

computed, but few researchers are aware of the chief danger with this type of model. Specifically, it is recommended that coefficients from the logistic regressions be put on the same metric (e.g., using a macro by Herr, 2016) as those coefficients obtained by OLS regressions in other parts of the mediation model. Researchers sometimes will do OLS analyses with categorical DVs and use the unstandardized B slopes and standard errors from these analyses to derive the mediation result, but this method is incorrect. Users must compute logistic regression where appropriate, and then typically convert the Bs and standard errors to the appropriate metric; Herr's Excel macro provides this function.

1.3. The Dataset

Many (perhaps a majority of) researchers use single-occasion (also known as *concurrent* or *cross-sectional*) datasets for mediation analyses. Because the authors should have conceptualized their mediation model as a temporal prediction (or a causal argument), then it is appropriate to examine such a prediction with experimental or longitudinal data. Although the seminal article by Baron and Kenny (1986) and subsequent work by Kenny (2014) make the point that a mediational model is an explicit causal argument (i.e., X leads to Z, which, in turn, subsequently leads to Y), researchers continue to propose and test models based on single-occasion or concurrent data. MacKinnon (2008), Jose (2013a), and Hayes (2013) all make the point that mediation with data collected at one point in time will not illuminate causal or temporal relations among variables. Instead, models based on data of this type only illustrate degrees of shared and unique variance among the variables. Although authors may argue that mediation based on concurrent data is suggestive of eventual findings with longitudinal data, some authors (Cole & Maxwell, 2003) are skeptical of this claim. Increasingly, reviewers and editors are requiring authors to use longitudinal or experimental data for mediation models.

1.4. Data Analysis

The simplest mediation model, shown in Figure 18.1, is composed of three variables: the independent variable (X), the mediating variable (Z), and the dependent variable (Y). Each path in the model has a name (given below). The *basic relationship* is the X to Y relationship by itself (often termed the *total effect* or the *c path*), in other words, a raw correlation or regression between the X and Y variables. The point of a mediation analysis is to try to explain at least a portion of the basic relationship by the inclusion of the third variable, Z, in the fashion depicted above. Once Z is added, the original *c* path is changed (typically reduced in strength), and it is now known as the *c'* (*c prime*) path (often termed the *direct effect*).

The method of data analysis should derive an estimate of the mediated or indirect effect that is sensitive, accurate, and unbiased. The indirect effect is typically estimated in one of three fashions: (1) using the “causal steps” approach described by Baron and Kenny (1986); (2) noting the size of the reduction from *c* to *c'*; or (3) multiplying *a* times *b*. The “causal steps” approach has been determined to be a poor method for determining significance (MacKinnon, Fairchild, & Fritz, 2007), and is no longer recommended as an inferential technique. MacKinnon, Lockwood, Hoffman, West, and Sheets (2002) compared a number of significance testing methods, and noted that both of the remaining methods ($c - c'$ and $a \times b$) have the limitation of suffering from low power unless appropriate methods are used to obtain the standard error estimate. Most researchers today use the $a \times b$ method, described in the Baron and Kenny article, with statistical significance determined with the use of Sobel’s test (Sobel, 1982).

However, because the product of *a* times *b* usually yields an asymmetrical distribution, inferential testing with methods (like Sobel’s test) that assume normal distributions lead to biased results.

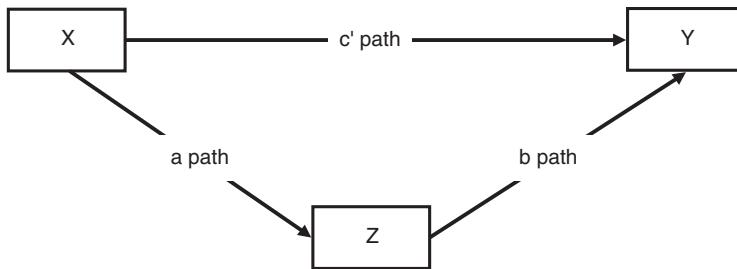


Figure 18.1 The Simplest Mediation Model.

With the advent of intensive computing algorithms, most mediation scholars recommend that bootstrapping be used instead of OLS regressions to compensate for lower power and asymmetrical distributions (Hayes, 2013; MacKinnon, 2008; MacKinnon et al., 2007). SPSS has recently added bootstrapping as a menu item for most of its analyses, and most SEM programs, such as AMOS, Mplus, and EQS, offer an option for bootstrapping estimates of indirect effects. PROCESS, an SPSS and SAS macro written by Hayes (2013), allows for the option of bootstrapping; packages in R (e.g., RMediation) have also started to emerge as well.

1.5. Statistical Output

The statistical outputs of a mediation analysis should be appropriate and clear. Reports that use Sobel's test typically report a z -statistic and its associated p -value (see Preacher's online applet for the calculation for the Sobel test (Preacher & Leonardelli, 2015). As noted above, key figures in the mediation field have recommended moving away from using the Sobel test, and instead recommend bootstrapping computation. Instead of a z -statistic and p -value, researchers are also recommended to report statistical significance with a confidence interval (CI) as it provides useful information about the size of the selected confidence interval (Jose, 2013a; MacKinnon, 2008). Bias-corrected confidence intervals are specifically recommended as they compensate for the asymmetrical distribution of $a \times b$ computations. For example, the researcher should be encouraged to provide statistical output such as the following: "The indirect effect from mindfulness through gratitude to happiness was found to be statistically significant, standardized indirect effect = .05, standard error = .013, 95% CI = [.04, .13]." These four outputs tell the reader the size of the indirect effect, the amount of variability of this estimate, and the resulting lower and upper bounds of the confidence interval. Because the confidence interval does not include the value of zero, it is interpreted as statistically significant. Please refrain from reporting CI information such as, "95% CI = .04 – .13." The dash between the lower and upper bounds of the CI is ambiguous: some writers mean for it to signify "to," and others are trying to indicate a negative numerical value.

1.6. Statistical Assumptions

Statistical assumptions should be demonstrated to be plausible. The type of variable in each slot must conform to the stipulations described above in 1.2. The distributions of continuous variables should approximate a normal distribution (no excessive skewness or kurtosis) (see MacKinnon, 2008). MacKinnon et al. (2007) further suggest that the residuals in the two regressions used to estimate the a and b paths be independent. They also recommend that authors should test for whether the IV and MedV significantly interact to predict the DV—if so, then the estimate of the indirect effect is likely to be biased. The chief issue in mediation models is misspecification, that is, incorrect placement of variables in the three mediation slots. As numerous writers have noted (e.g., Jose,

2013a), in cases of concurrent datasets, selection of variables for the three slots will not be guided by temporality; instead, most researchers rely on previous literature and sometimes logic. When available, temporality (e.g., in a longitudinal dataset) or experimental control (e.g., use of an experimental or quasi-experimental design) should be used to defend the choice of variable placement.

Another critical issue is the sample size. Fritz and MacKinnon (2007) conducted a simulation study to determine adequate sample sizes for six different types of mediation tests in order to achieve an empirical level of .80 power. For example, medium sized a and b path coefficients analyzed with a Sobel test need a minimum sample size of 90, whereas a bias-corrected bootstrap analysis requires a minimum sample size of 71 to achieve a level of .80 power.

1.7. Effect Size

Effect size estimates are unfortunately rarely computed and reported in research reports featuring statistical mediation. MacKinnon (2008) has described three different but mathematically related effect sizes, and Preacher and Kelley (2011) described two others as well. These techniques feature ratios or comparisons of the indirect effect with either the total effect or direct effect. The goal is to make a relative comparison of the indirect effect to either the total or direct effect. Jose's on-line applet MedGraph (Jose, 2013b) computes an *index ratio*, comparing the indirect effect to the total effect as stipulated by MacKinnon. For example, if the total effect is found to be .43, and the indirect effect ($a \times b$) is determined to be .11, a simple ratio of .11/.43 yields a ratio index of .26, suggesting that the mediating variable explained about one fourth of the total influence of the IV on the DV. This type of information, in conjunction with the previously identified statistical outputs (i.e., size of indirect effect, standard error, and 95% CI), contextualizes the finding and helps the reader understand the relative strength of the mediated effect. Be aware, however, that most effect size estimates do not compensate for sample size, so it is possible to derive a very small effect size (e.g., an index ratio of .05) for a statistically significant indirect effect—these usually occur with large samples, such as Ns greater than 500.

1.8. Interpretation

The chief problem with interpreting a mediation result is that researchers are too willing to ascribe causal relations among variables in studies featuring a concurrent dataset. Mediation scholars (see Hayes, 2013; Jose, 2013a; MacKinnon, 2008; Selig & Preacher, 2009) have consistently warned against this practice, but many writers still persist in using words like "cause," "influence," "affect," and related words. Given Kenny's (2014) argument that mediation is a causal argument, it is understandable that many researchers wish to apply mediation statistical techniques to their concurrent data and make a strong claim about variables affecting each other. Jose (2013a) makes the point that concurrent mediation is essentially an investigation of shared and unique variance among three variables, and although some causal relations may be hiding among these correlations, it is usually very unclear whether they illuminate how these variables would be related to each other over time. Maxwell (Maxwell & Cole, 2007; Maxwell, Cole, & Mitchell, 2011) has long argued that conclusions drawn from concurrent mediation are tenuous for the longitudinal context. In this vein, Jose (2016) has shown how concurrent mediation results may or may not generalize to longitudinal data. In sum, researchers should be cautious in using causal language in describing mediation results in the case of using a concurrent dataset. Findings based on longitudinal datasets may be on safer ground to use words such as "affect" and "predict," whereas findings based on experimental methods can be described with words such as "caused" at least for the a path; the b path is still subject to the effect of possible omitted variable confounders.

2. Moderation

2.1. Statistical Interaction

Examination of a moderation effect is based on an understanding of what a statistical interaction tells one about groups of individuals in one's dataset. In other words, moderation is about *who*, namely which groups of individuals in the dataset, exhibit particular directions and strengths of relations between variables. Research studying moderation must make clear the anticipated nature of the variable relations expected for which specific groups of individuals, and why.

2.2. Differences between Mediation and Moderation

Use of moderation is based on a clear awareness of how it differs from the approach of mediation. Mediation is a statistical technique designed to illuminate *how*, for example, how do people move from practicing mindfulness to a sense of gratitude, and then from the sense of gratitude to a state of subjective happiness? As noted above, a mediation finding tells us about the mechanism of moving from the IV, through the MedV to the DV. Moderation, in contrast, is not an argument about process; it is an argument about who displays what types of relations between the IV and the DV.

Considerable confusion still exists about the distinctions between mediation and moderation despite several decades of articles and books on this issue. Similarities are: (1) both can involve 3 or more variables; (2) both can be computed with ordinary regression techniques (as well as more sophisticated statistical techniques such as maximum likelihood and bootstrapping); and (3) both concern how a third variable is involved with a basic relationship between the IV and the DV. The chief difference is that moderation necessarily involves an interaction (or product) term between the IV and the third variable (termed the moderating variable or ModV). Mediation does not (typically or necessarily) involve an interaction term.

Contributing to confusion about the distinctiveness of these two approaches is the upsurge of interest in hybrid models that incorporate both mediation and moderation (see Desideratum 3.0 below).

2.3. Types of Variables

The types of variables should be appropriate for their respective roles in the model. Classic regression-based moderation involves a continuous IV, either a continuous or categorical ModV, and a continuous DV. Choosing the right type of variables for these three slots causes confusion for some researchers. In particular, researchers want to know whether the IV can be categorical. If the IV is dichotomous categorical (e.g., gender or experimental conditions), then it is more typical (although statistically equivalent) to run the statistical analysis as an analysis of variance (ANOVA). Accordingly, one may occasionally encounter the situation of a dichotomous categorical IV, a continuous ModV, and a continuous DV. In these cases, some researchers choose to perform a median-split on the ModV in order to make the variable amenable for ANOVA analysis; doing so, however, can come with a potential reduction in statistical power as well as diminished potential for interpreting the nature of the moderation. Categorical DVs require logistic analyses, and although efforts have been made to systematize so-called *logistic moderation* (Gelman & Hill, 2007; Hayes & Matthes, 2009), this approach is less frequently attempted in the social and behavioral sciences.

Most categorical ModVs are dichotomous (e.g., asthmatic vs. healthy), although Aiken and West (1991) have shown how ModVs that are composed of multiple categories can be dummy-coded and used in a moderation analysis. For example, ethnic group membership may be constituted by four categories: European American, African American, Hispanic American, and Other. Ethnicity would be dummy-coded, yielding three dummy variables, using one group

(e.g., European American) as the reference group. These three dummy variables would be multiplied individually with the IV, and then entered in the regression analysis: the IV, the three dummy codes, and the three interaction terms. Statistically significant interaction terms (one, two, or three) would be graphed individually to enable interpretation.

2.4. The Interaction Term

Preparation of the interaction term is controversial. Aiken and West (1991), in their seminal description of statistical moderation, recommended centering the IV and ModV before creating the product term, chiefly because this method expedites the hand-computation of the complex algebraic equations needed to plot the moderation graphs. This advice has been elevated to the status of being “required” of all moderations now, and many researchers routinely perform this data manipulation. Kromrey and Foster-Johnson (1998) have persuasively pointed out that this procedure is unnecessary for point estimates, and one can obtain an equivalent result for the interaction term by simply multiplying the two raw variables to create the product term. As Jose (2013a, p. 159) has described the situation, “centering changes only the intercept (and the size of the conditional main effects) and exerts no influence on the actual shape of the moderation result.” In sum, moderation findings derived from either method are equally valid. If the researcher wishes to make conclusions about conditional main effects, then centering may have some utility.

It should be understood that high correlations between the product term and the two constituent main effects are usually obtained (Aiken & West, 1991), and this degree of collinearity often serves to attenuate the power of the product term in typical regression-based analyses (Cohen, Cohen, West, & Aiken, 2003). Failure to replicate moderation results from other studies may be due to this bias.

2.5. Statistical Assumptions

The moderation analysis should be performed optimally and without bias. As moderation models are typically analyzed within a regression framework, typical regression assumptions apply (see Chapter 23); also, categorical variables (e.g., ModVs) should not exhibit excessive asymmetry in frequencies.

2.6. Graphing the Result

Graphing the moderation result is performed in such a fashion as to correctly and clearly depict the statistical finding. It is a truism of moderation that one must graph the result before interpretation can occur. Again, Aiken and West (1991) have laid down the definitive guidelines for properly and accurately graphing a moderation result. At the time that they wrote their book, most researchers were obligated to hand-compute cell means in order to generate a graphical depiction of the moderation. Today we have multiple on-line apps that perform this function more quickly and accurately than the laborious hand-computation. The applet of Preacher, Curran, and Bauer (2006a), linked with their journal article (Preacher, Curran, and Bauer, 2006b), is popular and provides useful additional outputs beyond the graph itself. Jose’s ModGraph is another commonly used graphing facility (Jose, 2013c), but there are many others as well.

2.7. Simple Slopes

To assist with the interpretation of a moderation result, many researchers also compute *simple slopes* for the lines of their graph. Aiken and West (1991) demonstrated the usefulness of

computing these values for graphs, and many researchers understand the value of including them in the paper as they assist the researcher in interpreting their result. A statistically non-significant slope is one that approximates a flat line, whereas significant slopes are understood to display either a significant positive or negative association between the IV and the DV. The simple slopes can be computed by hand from statistical output (see examples in Jose, 2013a), but, in addition, a number of graphing facilities also compute these values (see ModGraph by Jose, 2013c).

Following Aiken and West's (1991) suggestions, most researchers create figures with two lines for a dichotomous categorical moderator and three lines for a continuous moderator (i.e., low, medium, and high that are based on $-1SD$, the mean, and $+1SD$ respectively of the moderator). Some researchers choose to display only the $-1SD$ and $+1SD$ slopes, although this depiction is not optimal. In any case, the researcher should compute a simple slope for each line included in their graph.

2.8. Interpretation

Interpretation of the moderation result should be clear and accessible, and relevant to the proposed hypothesis. In the first instance, if the researcher proposed a moderation hypothesis, it should be directional. For continuous moderator variables, moderation results are typically classed into one of two groups: buffers or exacerbators. *Buffers* are moderators that yield a weaker slope for the high moderator condition (see Figure 18.2, top). According to the social support buffering hypothesis, individuals who report high levels of social support should manifest a weaker relationship between stress and depression. The figure shows that, in agreement with the hypothesis, the strongest slope was obtained for individuals who reported low social support, and the weakest slope was obtained for individuals who reported high social support. In contrast, an *exacerbator* yields the opposite pattern: in this case, the slope for high catastrophizing individuals is steeper than the slope for low catastrophizing individuals (see Figure 18.2, bottom).

Thus, when researchers pose a moderation hypothesis in the Introduction of their manuscript, they should say whether they expected the moderating variable to buffer or exacerbate the relation between the IV and the DV. To merely say something like “it was expected that hope would moderate the relation between positive events and happiness” is too vague and ambiguous. Instead, they should say something like, “it was expected that individuals reporting high levels of hope would evidence a stronger relation between positive events and subjective happiness than those individuals reporting low levels of hope.”

Moderation interpretations largely fall into two large categories: commenting on slopes of the lines, or commenting on outcomes (levels of the DV) for combinations of IV and ModV. The first approach has already been described above, and it is the most common technique used. An example of the second method (for Figure 18.2) would be something like, “The most depressed individuals in the sample reported both high stress intensity and high catastrophizing. Catastrophizing made little difference in depression scores for individuals reporting low stress levels, but it exerted a much stronger influence on individuals reporting high stress levels.” This second approach typically results in awkward, asymmetrical, and incomplete accounts of the moderation result (Jose, 2013a), and should generally be avoided.

The best interpretation of a moderation result combines all three variables and succinctly links them without proposing a causal relationship among them (unless an experimental manipulation is involved). For instance, in the example presented above, one might say, “Catastrophizing functioned as an exacerbator for the relation between stress intensity and depression. Individuals reporting the highest levels of catastrophizing yielded the steepest positive slope between stress intensity and depression.”

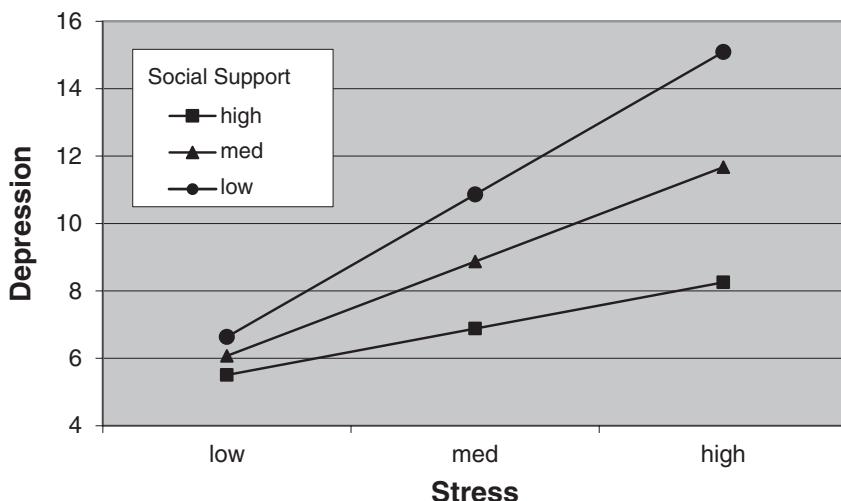
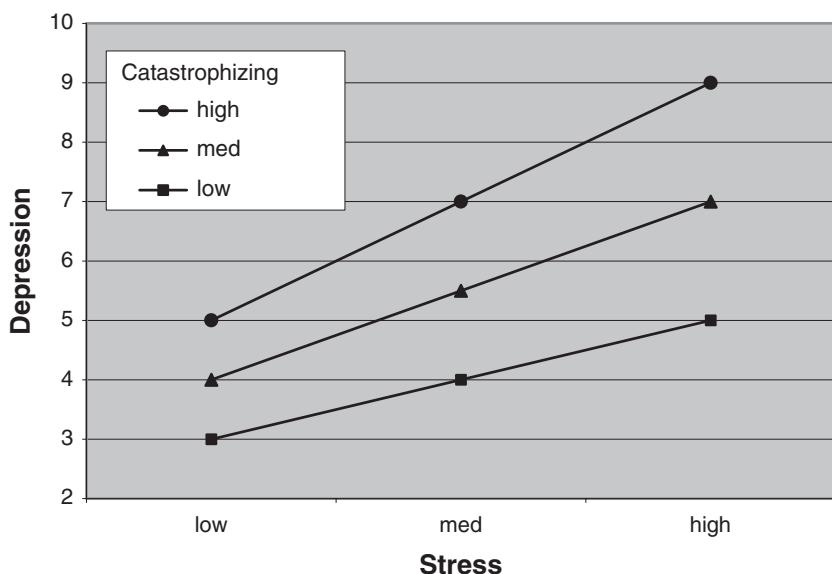
BUFFER:**EXACERBATOR:**

Figure 18.2 Moderating Variables—Buffers and Exacerbators.

It is also important to realize that in most moderation analyses, researchers use concurrent datasets, and consequently, the relations among all three variables are not appropriately described with causal language. For example, in the previous example, because all three variables were measured at a single point in time, one cannot argue conclusively that catastrophizing *caused* weaker or stronger relations between the IV and DV. Remember that the lines in the moderation figure depict groups of individuals who merely evidenced certain directions and degrees of association between the IV and DV. However, a few researchers consider moderations within experimental or longitudinal datasets, and in these cases, language can reflect temporality and/or causality.

3. Process Analysis

3.1. Process Analysis Model

When a researcher wishes to combine mediation with moderation in a single model, the appropriate process analysis is performed. Although researchers have long been interested in combining these two approaches in analyses (see examples of mediated moderation and moderated mediation in the seminal article by Baron and Kenny, 1986), the unfolding of a clear vision of how these should be done has been long in coming. An example of each of these two approaches will be presented first, and then I will describe the state-of-the-art treatment of these two designs last.

Figure 18.3 depicts a *mediated moderation* example similar to the one laid out by Baron and Kenny. Along the left side of the figure we see the IV (rumination), the continuous ModV (perceived control), and the resulting product term (rumination \times control). Their ability to predict levels of depression are both direct (i.e., direct paths from these three variables to depression) as well as indirect through the continuous mediating variable of anxiety. Thus, this figure depicts mediated moderation in that anxiety potentially mediates the influence of the rumination \times control interaction term on depression (see Jose, 2013a, pp. 249–253 for more information).

Few researchers pose a mediated moderation hypothesis because they are difficult to succinctly and clearly phrase. In this example, it would be something like, “the tendency for perceived control to buffer the relation between rumination and maladaptation is mediated by anxiety on depression.” As can be seen from this example, accessible interpretations are difficult to articulate, although moderation scholars have tried (Edwards & Lambert, 2007; Muller, Judd, & Yzerbyt, 2005). Hayes (2013, p. 358) argues provocatively that a mediated moderation “is really nothing other than a mediation analysis with the product of two variables serving as the causal agent of focus,” so he subsumes this approach within the broader category of *conditional process analysis*.

The other approach, namely *moderated mediation*, asks the question as to whether different groups of individuals exhibit the same or different degree of an indirect effect. In the following example an indirect effect from stress through resilient coping to negative adjustment was examined for younger and older children. As can be seen in Figure 18.4, the *a* path proved to be marginally significant and the *b* path proved to be non-significant for the younger group, and this group evidenced a non-significant mediation result, whereas the older group evidenced a significant mediation result. This variant is much more easily pitched as a hypothesis, and empirical testing is typically easier

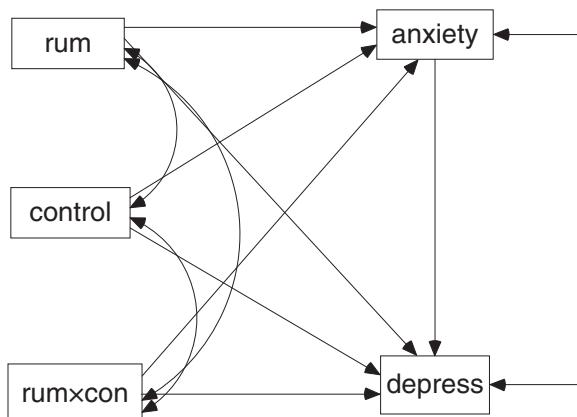


Figure 18.3 Mediated Moderation.

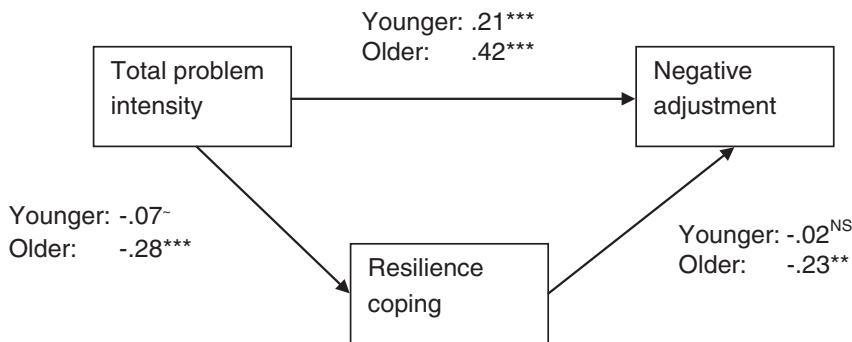


Figure 18.4 Moderated Mediation.

than mediated moderation too. A good interpretation of the moderated mediation result would be something like, “The moderated mediation analysis showed that a significant indirect effect was obtained for older children, indirect effect = .12, standard error = .02, 95% CI = [.07, .22].”

3.2. Types of Variables

The types of variables should be appropriate for the model proposed. The rules stipulated above concerning mediation and moderation apply here as well.

3.3. Statistical Analysis

It is possible to conduct both of these types of analyses in ordinary least squares (OLS) within regression analyses, and, in fact, many researchers use the PROCESS macro (for either SPSS or SAS) written by Hayes to do these analyses. PROCESS is a free downloadable macro, accompanied by Hayes’s book, which is entitled *Introduction to mediation, moderation, and conditional process analysis* (Hayes, 2013). The macro makes it possible to quickly and easily select variables in either SPSS or SAS datafiles and then to compute basic moderation or mediation models or more sophisticated models that combine the two approaches. Further, the macro takes advantage of the numerous options that both programs offer, so it is able to bootstrap estimates of indirect effects, generate confidence intervals, and create informative graphs of moderation results.

At the same time, many researchers rely on structural equation modeling packages such as AMOS, Mplus, lavaan (in R), and EQS to perform these analyses. The chief advantages of the SEM approach is that one can create latent variables, construct models involving multiple mediators and/or moderators, and obtain model fit indices more easily in SEM than with regression-based programs (see Chapter 33).

3.4. Post-Hoc Probing

As noted above, post-hoc probing of moderation through simple slopes should be performed correctly. Further (and PROCESS does a good job of this), post-hoc probing of moderation groups in moderated mediation should be performed. For example, if a researcher obtains a significant moderation of a mediation, it is important to compare and contrast the indirect effects for the different moderation groups (either categorical or continuous).

3.5. Interpretation

As noted above, explanation of a mediated moderation model is not simple, but through discussion of the beta weights and graphed moderation results, one should be able to articulate what the finding suggests. And with the proper use of simple slopes, one should be able to pose an informative interpretation of a moderated mediation results.

References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. A. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, 112, 558–577.
- Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods*, 12, 1–22.
- Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18(3), 233–239.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
- Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, 76, 408–420.
- Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis*. New York: Guilford Press.
- Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior Research Methods*, 41, 924–936.
- Herr, N. (2016). Mediation with dichotomous outcomes. Retrieved May 22, 2016 from www.nrhpysch.com/mediation/logmed.html.
- Jose, P. E. (2013a). *Doing statistical mediation and moderation*. New York: Guilford Press.
- Jose, P. E. (2013b). *MedGraph: An on-line programme to graphically depict mediation among three variables: Version 3.0*. Wellington, New Zealand: Victoria University of Wellington. Retrieved May 31, 2016 from <http://pavlov.psyc.vuw.ac.nz/paul-jose/medgraph/medgraph.php>.
- Jose, P. E. (2013c). *ModGraph-I: A programme to compute cell means for the graphical display of moderational analyses: The internet version, version 3.0*. Wellington, New Zealand: Victoria University of Wellington. Retrieved May 30, 2016 from <http://pavlov.psyc.vuw.ac.nz/paul-jose/modgraph>.
- Jose, P. E. (2016). The merits of longitudinal mediation. *Educational Psychologist*, 51(3–4), 331–341. DOI: 10.1080/00461520.2016.1207175
- Kenny, D. A. (2008). Reflections on mediation. *Organizational Research Methods*, 11, 353–358.
- Kenny, D. A. (2014). *Mediation*. Retrieved 5 November, 2015 from <http://davidakenny.net/cm/mediate.htm>.
- Kromrey, J. D., & Foster-Johnson, L. (1998). Mean centering in moderated multiple regression: Much ado about nothing. *Educational and Psychological Measurement*, 58, 42–67.
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Lawrence Erlbaum and Associates.
- MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, 58, 593–614.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83–104.
- Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, 12, 23–44.
- Maxwell, S. E., Cole, D. A., & Mitchell, M. A. (2011). Bias in cross-sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behavioral Research*, 46, 816–841.
- Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When mediation is moderated and moderation is mediated. *Journal of Personality and Social Psychology*, 89, 852–863.
- Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006a). Probing interactions in multiple linear regression, latent curve analysis, and hierarchical linear modelling. Retrieved from <http://quantpsy.org/interact/index.htm>.
- Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006b). Computational tools for probing interaction effects in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, 31, 437–448.
- Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, 16, 93–115.
- Preacher, K. J., & Leonardelli, G. J. (2015) *Calculation for the Sobel test: An interactive calculation tool for mediation tests*. Retrieved 11 November, 2015 from <http://quantpsy.org/sobel/sobel.htm>.
- Selig, J. P., & Preacher, K. J. (2009). Mediation models for longitudinal data in developmental research. *Research in Human Development*, 6, 144–164.
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhart (Ed.), *Sociological methodology 1982* (pp. 290–312). San Francisco, CA: Jossey-Bass.

19

Meta-analysis

S. Natasha Beretvas

Meta-analysis entails a set of analytical techniques designed to synthesize findings from studies investigating similar research questions. While meta-analysis includes narrative integration of results, the current chapter will focus only on quantitative meta-analysis. Meta-analysis permits summary of studies' results and is designed for scenarios in which the primary studies' raw data are not available. The meta-analytic process involves summarizing the results of each study using an effect size (ES), calculating an overall average across studies of the resulting ESs, and exploring study- and sample-related sources of possible heterogeneity in the ESs. The overall average ES provides a single best estimate of the overall effect of interest to the meta-analyst. Meta-analysis can be used to explore possible differences in ESs as a function of study and sample characteristics. In the seminal article in which the term *meta-analysis* was coined, Smith and Glass (1977) used meta-analysis to summarize results from studies that had assessed the effectiveness of psychotherapy. Thus, treatment effectiveness results provided the first type of ES to be synthesized using meta-analysis. Since the 1970s, the field of meta-analysis has grown to include methods for conducting the synthesis of other types of ESs including correlations, transformations of odds-ratios, validity coefficients, reliability coefficients, and so forth.

Many textbooks provide detailed descriptions of the meta-analytic process. Texts by Lipsey and Wilson (2001), Rosenthal (1991), Card (2012), and Borenstein, Hedges, Higgins, and Rothstein (2009) provide excellent introductions to meta-analysis. Hunter and Schmidt's (1990) textbook provides the seminal resource for meta-analysts interested in correcting ESs for artifacts (see Desideratum 11). Books by Cooper, Hedges, and Valentine (2009) and Hedges and Olkin (1985) are recommended for readers with more technical expertise. Meta-analysts interested in a text devoted to description of ways to assess and correct for publication bias should refer to Rothstein, Sutton, and Borenstein (2005). Desiderata for studies that involve use of meta-analysis are contained in Table 19.1 and thereafter they are discussed in further detail.

1. Theoretical Framework and Narrative Synthesis

As with any manuscript, a summary of past research must justify the selection of the study's research question. Similarly, a meta-analysis must be prefaced by a narrative synthesis summarizing results found in previous studies that are to be integrated in the meta-analysis. The narrative synthesis must clarify the specific research question associated with the effect size (ES) that is being synthesized.

Table 19.1 Desiderata for Meta-analysis.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. A theoretical framework is provided that supports the investigation of the effect size (ES) of interest and includes a narrative synthesis of previous findings.	I
2. Type of ES of interest in the study is specifically detailed (e.g., correlation, standardized mean difference).	I, M
3. Databases searched and keywords used to find relevant studies are listed, as well as criteria for deciding whether to include a study in the meta-analysis.	M
4. Formulae used to calculate ESs are provided or referenced, and any transformations used (e.g., to normalize or stabilize ES sampling distributions) are made explicit.	M
5. The coding that is used to categorize study and sample descriptors is provided.	M
6. Estimates are provided that describe the interrater reliability of the information coded in each study.	M, R
7. If study quality is assessed, a description is provided detailing how it is assessed and how study quality is incorporated into the meta-analysis.	M
8. For weighted analyses, the type of weights used is provided.	M
9. Methods used to handle within-study ES dependence (e.g., multiple ESs per study) are described.	M
10. Methods used to access, assess, and handle missing data are detailed.	M
11. If relevant, the method used to correct for artifacts is described.	M, R
12. Homogeneity of ESs is assessed.	M, R
13. Statistics describing the resulting meta-analytic dataset that was gathered and including pooled estimates of the effect size of interest are provided along with associated standard errors (and/or confidence intervals).	R
14. Inferential statistics describing the relation between the study and sample descriptors and the effect size are presented.	R
15. Interpretation is offered describing the practical significance of the ES magnitude and direction and the relation between moderators and the ES.	D

* I = Introduction, M = Method, R = Results, D = Discussion.

The narrative synthesis summarizes in words what previous research has found in terms of the patterns of results relevant to the ES of interest. While many studies might investigate the same basic research question, the studies can be distinguished by various sample and study composition descriptors. Examples of descriptors include demographic variables such as gender, ethnicity and age, and characteristics of the study's design such as the type and duration of an intervention, the outcome measure used, the research context, and the experimental design used. The review of previous literature should clarify and identify the importance and relevance of these descriptors to the ES of interest. This then lays the groundwork for investigation of relations between these descriptors (termed *moderators*) and the ES in the ensuing meta-analysis.

2. Effect Size

The fundamental unit of any meta-analysis is the effect size (ES). An ES provides a parsimonious descriptor containing information about the direction and magnitude of the results of a study. The most commonly used meta-analytic ESs include the standardized mean difference, the correlation (representing the relation between two variables), and the odds ratio. A meta-analytic ES describes the relation between a pair of variables. Operationalization of each of the two variables should be clarified and justified. For example, if student achievement is one of the pair of variables of interest in a meta-analysis then the sorts of test scores that qualify as student achievement should be clarified. Description

of the research question of interest in the meta-analysis should clarify the ES being investigated both in terms of the statistical type as well as the operationalization of each of the relevant two variables.

3. Study Inclusion Criteria

The Introduction (see Desiderata 1 and 2) should have clarified the components necessary for deciding to include a study's results in the meta-analysis. A section in the Methods section of a meta-analysis must detail how the relevant studies and results were found. The databases (e.g., PsycInfo, ERIC) that were searched, the types of studies (e.g., peer-reviewed publications, dissertations, conference presentations) and the keywords used must be identified. Any additional means used for finding relevant studies that were not initially identified should also be described (e.g., using the References section in studies that had been identified in the database search, contacting study authors, etc.). In addition to emphasizing the acceptable operationalizations of the constructs relevant to the ES, the population of interest should be described. For example, a researcher might solely be interested in an ES for adults and thus data would be excluded from any study that had investigated the relevant variables for adolescent respondents.

Meta-analysts must also decide on the types of study designs that qualify for inclusion. Some meta-analysts include only results from studies employing purely experimental designs while others also include quasi-experimental studies' results. Some studies necessitate the use of single-subject designs for which there is still controversy in terms of how to meta-analytically synthesize the results. If a more general inclusion strategy is used, then meta-analysts should code the relevant design features and summarize descriptively or inferentially the potential differences in resulting ESs (see Desiderata 13 and 14).

4. Calculation of Effect Sizes

The (statistical) type of ES being synthesized should have been clarified in the Introduction (see Desideratum 2). Results reported in the primary studies being synthesized are not all in the same format. For example, a meta-analyst might be interested in synthesizing a treatment's effectiveness using a summary of standardized mean differences across studies. Some studies might provide the treatment and control groups' means and standard deviations for the relevant outcome. Other studies might instead provide the results of an independent samples *t*-test comparing the treatment and control groups on the outcome score. Results in both formats can be converted into a standardized mean difference ES metric. Authors should clarify any conversion formulas they use to convert studies' results into a common ES metric.

In addition, some estimators of the most commonly used ES (the standardized mean difference) have been found to be biased. There are a number of ways that this ES is calculated (including, most commonly, Cohen's *d*, Glass's Δ , and Hedges's *g*). The meta-analyst must clarify and justify which estimate of the standardized mean difference is being used.

Sampling distributions of most of the typical untransformed ESs (e.g., standardized mean difference, correlation, odds ratio) have been found to be non-normal. One of the purposes of quantitative meta-analysis is to use statistical tests of the ES and its relation with sample and study descriptors. Thus, it is important to use the transformations that normalize (and stabilize the variances) of the sampling distributions of these ESs. Meta-analysts should detail the formulas that are used to transform the resulting ESs estimates for ensuing statistical analyses.

5. Coding of Study and Sample Descriptors

A host of variables typically distinguish the studies and samples being synthesized and might be related to the resulting ESs. Sources of the possible heterogeneity in ESs across studies can and should

be explored using these variables. When gathering primary study data to be used for calculating the ESs, meta-analysts should also gather information associated with the samples in each study. Sample size is an essential variable that must be coded as it provides information about the precision of each study's ES and can be used as a weight in resulting ES analyses (see Desideratum 8). Demographic information (such as age, gender, and ethnicity composition of the sample) can also be coded and used in the meta-analysis. Characteristics of each of the two variables whose relation is being synthesized should also be coded and captured. For example, in a study summarizing a family-based treatment's effectiveness in reducing internalizing disorders, the meta-analyst might have multiple constructs such as depression and anxiety that qualify as internalizing disorders. Each type of outcome could be coded to explore possible differences in the treatment's effectiveness for the more specific kinds of internalizing disorders. This can lead to multiple ES estimates being gathered per study and thus some dependence that must be handled (see Desideratum 9). There might also be characteristics of the implementation of the treatment that distinguish the primary studies and define the resulting ESs. In the current internalizing disorders example, interventions might be designed to involve both parents and children or they might be designed only for parents. Thus, categories distinguishing interventions could also be coded and collected. In addition, and specifically for intervention effectiveness meta-analyses, some studies might report results for more than one intervention. As with a study reporting multiple outcomes, the dependence resulting from multiple ESs per study needs to be appropriately handled (see Desideratum 9). Last, for a meta-analysis of intervention effects, not all studies might compare all interventions of interest. Network meta-analysis procedures can be used to synthesize results from studies comparing different sets of interventions' effects with each other.

Facets of a study's design can also be gathered and included in the meta-analysis (as described in Desideratum 14). As mentioned in Desideratum 3, a study's design should be coded as it can later be used to explore potential differences in ESs resulting from differing experimental designs. Some meta-analysts code "study quality" and evaluate its relation to the ES values. Some meta-analysts correct their ESs for artifacts to match what the ESs would be for a perfect study that used an infinitely large sample with access to perfectly reliable and valid test scores. If interested in correcting for artifacts, the meta-analyst would gather relevant information including, for example, the reliability of scores on the measures of interest (see Desideratum 11). Additional selection of study and sample descriptors should be founded in the meta-analysts' research questions in terms of what they hypothesize might explain variability in ESs.

Values for some of the descriptors might differ for samples within a study. Group sample size in a meta-analysis of a treatment's effectiveness (i.e., using the standardized mean difference ES to summarize the difference in means between a treatment and control group) provides a simple example of a sample-level descriptor. Other descriptors might only vary across studies (e.g., whether the population being assessed was college students). The coder must clarify the distinction between such sample-level descriptors and study-level descriptors that differ across, but not within, studies. This information is essential to inform selection of the analytic technique that best matches the data's structure and for addressing how to mean-center moderators in meta-regression (moderator) analyses.

One last piece of information about coding must also be provided. Unfortunately, the information sought by meta-analysts is not always presented in the primary studies. It is important for meta-analysts to clarify how they attempted to gather moderator variable values that are missing as well as to detail the methods used to handle the missingness (see Desideratum 10).

6. Interrater Reliability

Given the amount of information that needs to be gathered and coded in a meta-analysis, it is typical to involve at least a couple of researchers as coders. It is thus important to provide a description of the reliability of the coding that was conducted. If data indicate that coding is not reliable, further

coding training should be conducted and consensus about each study's codes must be reached. At the very least, the average (median) percent agreement for each variable should be reported in the meta-analysis. Use of kappa, weighted kappa, or the intraclass correlation to provide additional measures of interrater agreement is also encouraged (see Orwin & Vevea, 2009, for additional details; see also Chapter 10, this volume). While it would be optimal for at least two coders to code every study in the meta-analysis, that sometimes is not feasible. If this is the case, then at least a reasonable proportion of studies should be coded by at least two raters with sufficient justification provided for not having two raters code all studies. Given a lack of complete agreement in the coding that is done by the two raters, the meta-analyst must describe how differences were resolved and consensus reached through discussion and possible re-specification of codes used.

7. Study Quality

Since the introduction of the term *meta-analysis* in the 1970s (Glass, 1976), researchers have argued about how to handle differences in research designs' quality when synthesizing studies' results. Researchers agree that, at the outset, meta-analysts must select and justify a research design quality criterion for study inclusion. In addition, meta-analysts are encouraged to gather and code information (see Desideratum 5) on a study's design that might differentiate studies' ES results.

All sorts of factors might impact the quality of a study's design and thus also affect the ES results. Those factors include group selection and assignment, experimenter expectations (e.g., whether a study is blinded), psychometric properties of measures, and many more. It is up to the researcher to select the pool of possible design quality variables of relevance to the meta-analysis. Meta-analysts can use the resulting variables descriptively or use them as moderating variables in ensuing analyses (see Desiderata 13 and 14).

8. Weights

As with any consistent estimator, the precision of an ES estimate is greater when it is based on larger sample sizes. Thus, when pooling ES estimates, meta-analysts typically weight ESs by some function of their associated sample sizes (see Desideratum 13). When testing meta-regression models (see Desideratum 14) designed to explore the variability in ESs using study and sample descriptors as moderators, meta-analysts frequently estimate models involving these same N -based weights. In either scenario, more weight is assigned to estimates based on larger sample sizes. The most commonly used weights are either the inverse of N or the inverse of the variance of the ES of interest (which will also be a function of N). The weight entailing the inverse of the conditional variance results in the most efficient pooled estimate of the population ES and thus is recommended here. However, the meta-analyst should clarify the function of N that is being used as the weight.

9. Handling Dependent ESs

Studies can frequently contribute multiple ESs to a meta-analysis. These multiple ESs can be considered dependent if they are based on the same sample. For example, in meta-analyses designed to assess intervention effectiveness (i.e., comparing two groups on an outcome), a study can provide results from comparing the two groups on each of multiple, related outcomes. Given that sufficient data are provided in the study for each outcome that corresponds to the construct of meta-analytic focus (e.g., depression and anxiety might both qualify as internalizing outcomes), an ES can be calculated. The resulting two standardized mean difference ESs are assumed dependent because the ESs describe a common sample. Equally important is that the ESs are based on measures that are themselves correlated (e.g., depression and anxiety).

Alternatively, in a meta-analysis of the correlation between two variables, multiple dependent ESs would result from a study that provided correlation estimates between pairs of variables both of which matched the constructs of interest. This study would qualify as a *multiple-endpoint* study. For example, the meta-analyst might be interested in the correlation between internalizing disorders and academic achievement. If a study reports the correlation between, say, depression and SAT scores and the correlation between anxiety and SAT scores, then both correlations could be used to calculate ESs for later analysis. The dependence would again originate in the use of a common sample for estimation of the two ESs.

Another example of the source of possible dependence commonly found in meta-analyses of intervention research might originate in a study reporting results from comparing three groups on an outcome. This study would be an example of a *multiple-treatment* study. For example, a meta-analyst might be interested in summarizing the effectiveness of parental involvement interventions for improving internalizing disorders. A primary study might evaluate the internalizing disorders of three groups, two of which involve differing implementations of a parental involvement treatment and a control group. Two effect sizes could be calculated with one comparing the internalizing disorder scores of the first intervention group with the control group. The second ES would describe the difference in internalizing disorders between the second intervention group and control group. Given the involvement of the same control group in the calculation of the two ESs, the ESs would be considered dependent. Other possible dependencies might be encountered between pairs of effect sizes within a study that should also be handled appropriately.

Meta-analysts have a choice of methods they can use to handle effect size dependence. Some researchers choose to ignore the dependence which will negatively impact the validity of the associated statistical conclusions. Other meta-analysts might choose a single effect size to represent each study. For example, this “best” ES might be based on the measure with the best psychometric functioning in each study. Still others might calculate a weighted or simple average of each study’s multiple ESs and use the result as the single ES for each study. While use of a single ES per study (selected via aggregation or deletion of the study’s multiple ESs) does result in an analysis of independent ESs, it overly reduces the available database and thus possible information. In addition, this reduces ensuing statistical power. It also unnecessarily reduces the possible heterogeneity in the ESs.

Another option available for handling dependent ESs involves modeling the multivariate nature of the dataset. Several options are available with use of generalized least squares (GLS) estimation procedures being the most commonly used method. The primary problem with the use of multivariate modeling to handle possible dependencies is that additional data must be gathered from the primary studies. For example, to use GLS for synthesizing results from multiple-endpoint studies, meta-analysts must use values for the correlation among scores on the multiple endpoints. However, it is sometimes possible to impute reasonable values for this correlation and, despite their complexity, GLS methods have been found to work well for handling meta-analytic dependence in some scenarios. Two more recent methods suggested for handling within-study dependence include the use of robust variance estimation (Hedges, Tipton, & Johnson, 2010) and the multi-level meta-analysis model (van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013). Regardless of the method used, the meta-analyst must note the types of dependence that they encountered in their dataset. They must also describe and justify their choice of method used to handle this dependence.

10. Methods for Handling Missing Data

As with most social science datasets, analysis of meta-analytic datasets is also hampered by missingness. This can result from primary studies not reporting sufficient statistical information permitting

calculation of an effect size. Alternatively, primary studies might not have gathered or not reported all information of interest to the meta-analyst. For example, a meta-analyst might be interested in explaining heterogeneity in an ES using a variable representing the percent of participants who were female. Not every study will necessarily report the percent of participants who are female. Meta-analysts need to detail and justify how they handled missing data.

As with primary study analyses, there are a host of options that meta-analysts can use to handle missingness. There are similar caveats associated with these techniques when used in a meta-analytic context. For example, use of listwise or pairwise deletion still requires the assumption that data be missing completely at random and frequently results in large reductions in data available for a meta-analysis. These methods are not strongly recommended for use with meta-analytic data. Use of single-value imputation is not uncommon in meta-analysis (e.g., using a mean of reported values' information, or using a value that is reasonable based on patterns of values reported in other studies with similar participants). Single-value imputation can be recommended although its use inappropriately reduces the associated variability. Use of multiple imputation (MI) is still rare in meta-analysis but if it is used, then the missingness is assumed to be missing at random. Further methodological research is needed to assess the functioning of MI especially given the weighting typically used in meta-analysis, however, it would seem likely to function best as a method for handling missing meta-analytic data.

Meta-analysis is also criticized for another form of missingness peculiar to this technique, namely, missingness due to *publication bias*. Publication bias is a term that refers to the scenario where only studies with statistically significant results are reported ("published") and only studies that are published (i.e., available) can provide data that can be synthesized in a meta-analysis. Clearly this kind of missingness will bias resulting ESs. There are many different ways researchers use to assess whether publication bias might exist. Graphical displays, such as the funnel plot are sometimes used. ES estimates based on smaller sample sizes would be expected to vary more than for studies based on larger sample sizes although the average ES should not depend on sample size. Funnel plots involve graphing ES estimates against their associated sample sizes and provide a graphical way of assessing whether this pattern holds. If the plots are skewed, then this might be inferred as evidence of publication bias although other explanations are also possible.

Indices are also available to assess potential publication bias. The fail-safe number as well as modifications thereof, trim-and-fill estimates can also be used to evaluate the potential for publication bias. Last, some meta-analysts use inferential tests of publication bias (e.g., Begg's rank correlation test, Egger's regression, and funnel plot regression). The reader is strongly encouraged to refer to any of the meta-analytic texts (especially Cooper & Hedges, 1994, and Rothstein et al., 2005) to find out further details about these different procedures. Meta-analysts are encouraged to use multiple methods for assessing publication bias including at least the trim-and-fill method as well as one of the regression methods despite their limited statistical power.

Meta-analysts should try to contact the primary study authors to obtain information that might not have been reported. In the absence of this information and if evidence supports the possibility of publication bias, meta-analysts are encouraged to use any of the variety of methods available for correcting for publication bias. In particular, the trim-and-fill correction and the use of weighted distribution theory-based approaches are strongly recommended.

11. Correction for Artifacts

Some meta-analysts use artifact correction procedures to correct for artifactual errors resulting from imperfect research scenarios. These correction procedures are designed to correct resulting ES estimates so that they represent results under ideal research scenarios (for example, they can be used to

correct an ES estimate so that it represents the ES estimate based on perfectly reliable and valid test scores). The most commonly used correction is the correction for attenuation that can result from the lack of perfect reliability of scores on social science measures. Other corrections include correction for dichotomization of continuous variables and for restriction of range. Use of these procedures involves obtaining additional information (e.g., internal consistency reliability estimates for the relevant outcomes) to correct the relevant ES as well as its associated variance estimate. Use of artifact correction can also affect the sampling distributions assumed for the resulting ESs. Meta-analysts must specify which artifacts they might be correcting for and how. There is no consensus in the field about the use of these artifact correction procedures. Given the difficulties encountered in terms of gathering realistic values to calculate the corrections and their effects on the ESs' sampling distributions, the validity of the resulting corrections and of analyses conducted using the corrected ESs seems questionable.

12. Homogeneity of ESs

Meta-analysis is used to synthesize results from a multitude of studies designed to assess the same research question. While replication is encouraged in research, most studies do not exactly mimic each other. Studies tend to involve some subtle (or not so subtle) variation on a previous but similar study. Samples from different populations might be used (e.g., adults versus adolescents or college students, clinical versus non-clinical respondents, populations with different demographic information). Different implementations of an intervention might be tested. Different measures of a related but distinct construct might be investigated. This means that the resulting effect sizes might not come from a single population (sampling distribution of effect size estimates) with a single true effect size. Instead, it is more likely that while some of the variability in effect size estimates is due to sampling error, some of the variability is also attributable to random effects. In other words, the estimates do not come from a single population.

Meta-analysts should test the heterogeneity of the effect size estimates they gather. Methodological researchers have consistently supported use of the Q-test statistic designed to test the null hypothesis of homogeneous ESs. If the variability in the effect sizes is found to be more than could be solely attributed to sampling error, then this affects the model that should be assumed when conducting ensuing statistical analyses. Excess heterogeneity means that a random effects model should be assumed. If the effect sizes can be assumed homogeneous, then a fixed-effects model can be assumed. A meta-analytic researcher should clearly identify which model was assumed for all analyses including estimation of both pooled estimates as well as for analyses designed to investigate sources of variability in effect size estimates using the moderating variables detailed in Desideratum 5. Note that when reporting meta-regression model results, the choice made between a fixed- versus mixed-effects model should be clarified.

13. Descriptive Statistics

Meta-analysts should describe the resulting data that were gathered. This includes the availability of sample and study descriptors as well as information that could be used to calculate ESs. Some meta-analysts provide a table listing each study and associated descriptive information (such as the sample size underlying an ES as well as other study and sample descriptors as noted in Desideratum 5). This table usually also provides every ES or an overall ES for each study (see Desideratum 9). All meta-analysts present ES estimates pooled across studies for each outcome of interest and usually for levels of categorical moderating variables of interest. Along with all pooled estimates, associated standard error (and/or confidence interval) estimates should be provided. The (random-, mixed-, or fixed-effects) model that is assumed for these should already have been noted (see Desideratum 12).

14. Inferential Statistics

Results summarizing the tests of relation between moderators (see Desideratum 5) and the ES should be presented. Meta-analysts testing a number of moderating variables should consider use of (weighted) meta-regression model for testing the concurrent inter-relations. Conducting a multitude of statistical tests can lead to inflated Type I error for meta-analytic data as with any other kind of data. Controls such as the use of Bonferroni's correction to the nominal alpha level should be considered. Last, meta-analysts should appropriately model the meta-analytic data's structure. For example, in a meta-analysis involving multiple ES estimates per study, some of the moderators might be sample-level descriptors while others might be at the study level. The same considerations about centering of predictors (moderators) as those addressed with multilevel modeling of primary study data still apply.

15. Practical Significance

As with any empirical study, detection of statistical significance (or non-significance) should be interpreted within a context. While some researchers might cite rules of thumb for cutoffs representing small, moderate, and large effect sizes, interpretation of an effect size's magnitude should be made in the explicit context in which the effect size is calculated. For example, an ES estimate of 0.001 would qualify to most researchers as minuscule. However, in a test of aspirin for reducing heart attacks an ES estimate (R^2) of 0.011 was deemed sufficiently large that the trial was prematurely halted to stop "harming" placebo recipients who were not being given the aspirin (cited in Rosenthal, 1994). Thus, rules of thumb for describing an effect's size should be used with caution. Instead, the researcher should consider the magnitude and direction of the ES estimates in the context in which they are being assessed. Similarly, the strength (and direction) of the relation between the moderating variables and the ES should be interpreted at a practical rather than solely a statistical significance level.

References

- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. New York: Wiley.
- Card, N. A. (2012). *Applied meta-analysis for social science research*. New York: Guilford Press.
- Cooper, H. M., & Hedges, L. V. (Eds.) (1994). *The handbook of research synthesis*. New York: The Russell Sage Foundation.
- Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *Handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher*, 5, 3–8.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1, 39–65.
- Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Orwin, R. G., & Vevea, J. L. (2009). Evaluating coding decisions. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 177–203). New York: Russell Sage Foundation.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.
- Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.) (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Hoboken, NJ: John Wiley and Sons.
- Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32, 752–760.
- Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three level meta-analyses of dependent effect sizes. *Behavior Research Methods*, 45, 576–594.

20

Monte Carlo Simulation Methods

Daniel McNeish, Stephanie Lane, and Patrick Curran

For many applied researchers within the social and behavioral sciences, appropriate statistical and measurement tools are selected and employed to be able to make reliable and valid inferences about substantive hypotheses under study. For other researchers, however, the methods used to practice statistics and measurement and the determination of what are the most appropriate methods is a substantive area in and of itself. To these quantitative methodologists, the research questions of interest focus more on the properties and performance of the methods themselves rather than on the methods as applied to empirical data. When possible, these methodological questions are best addressed mathematically using, for example, analytical derivations. However, when evaluating and systematically manipulating conditions such as sample size, missing data, and distributional characteristics, these analytical deviations often become intractable and alternative approaches must be employed to study the method under examination. Monte Carlo computer simulation offers one such approach. In a typical application, data are simulated to be consistent with the model structure and/or assumptions underlying the quantitative method under study; models are then fitted to those simulated data and predefined outcome measures of interest are evaluated to gauge the method's performance. Although computer simulation methodology can offer unique and powerful insights into a broad class of quantitatively relevant research hypotheses, there exists tremendous variability in how these methods are used in practice. Sources of variability include theoretical motivation, experimental design, model selection, data generation, data analysis, and reporting of results, all of which can combine to make evaluating a Monte Carlo simulation study a challenging task. To help structure the many ways in which computer simulation might be used in practice, 12 specific characteristics are described that should typically be present in methodological research employing this investigative approach.

1. Motivating Research Question

As computational resources continue to become increasingly accessible, the barriers to conducting a Monte Carlo simulation are not as great as they once were. Indeed, some popular software programs have built-in syntax to allow researchers to conduct Monte Carlo simulation studies with minimal background in computer programming, including data generation, model fitting, and summary of the obtained results. While these lower access barriers are of course positive, they do allow simulation studies to be conducted by an increasingly broad community of researchers. Given the relative ease with which

Table 20.1 Desiderata for Monte Carlo Simulation Methods.

<i>Desideratum</i>	<i>Manuscript Section*</i>
1. A clear statement of the motivating research question is articulated.	I
2. A comprehensive review of existing analytic theory and prior empirical findings is presented.	I
3. A discrete set of theoretically derived research hypotheses to be empirically tested using computer simulation methods is formulated.	I
4. An unambiguous justification is provided supporting the use of computer simulation over other available methods, particularly analytic derivations.	I
5. A clear description of the manipulated independent variables and a rationale for the particular levels chosen are presented.	M
6. A clear definition and description are provided of all dependent variables to be empirically examined.	M
7. A clear description is presented of the model structures used to generate the simulated data.	M
8. A detailed articulation is provided of precisely what procedures were used to generate the simulated data. The generating code itself is provided as an appendix or made available to readers at an easily accessible location.	M
9. A clear description is provided of the exact model structures that were fitted to the simulated data.	M
10. A clear explanation is presented regarding the methods that were used to analyze the dependent variables under study.	M,R
11. A comprehensive yet lucid presentation of the empirical results is provided, often in both tabular and graphical forms.	R
12. A summary of key findings and the articulation of empirically informed inferences back to the underlying theoretical model are provided.	D

* I = Introduction, M = Method, R = Results, D = Discussion.

a simulation study can be conducted, we must think even more critically about the unique contribution of the chosen modeling framework and experimental manipulations relative to what is already known.

Therefore, at the onset of a study reporting Monte Carlo simulation results, the motivation should clearly communicate how the study results will help to move the methodological literature forward. For instance, a simulation study comparing the performance of different estimators would be relatively straightforward to conduct. In designing such a study, the motivation might be to inform empirical studies regarding best practice for data analysis under certain known conditions. For example, it might be the case that the finite sample performance of certain estimators has not been well studied. Alternatively, one estimator may make more rigid assumptions that are not generally tenable with real data; or perhaps a particular estimator has difficulty converging on a stable solution. Regardless, it is important that simulation methodology be thoughtfully used as a method of empirical data collection to test theoretically derived research hypotheses in precisely the same way as would be done in any other substantive field of study. The extent to which a simulation study is not being used to directly test a specific research hypothesis undermines both the internal and external validity of the obtained results.

2. Review of Previous Findings

Once the motivating research question has been delineated, a careful review of prior research, whether analytic or simulation-based, should be presented. Optimally, this review should provide a

sense of the scope of prior evaluative work on the topic of interest. Relevant characteristics of prior studies may include the experimental conditions that were manipulated (e.g., effect size, number of items) and the outcome measures that were evaluated (e.g., bias, efficiency). This review serves to contextualize the proposed simulation study, and helps to inform the conditions or levels manipulated in the study (see Desideratum 5). Related to Desideratum 1, a thorough review of previous research demonstrates that relevant gaps in the literature exist and can pinpoint precisely where those gaps occur. The ultimate purpose of the review is to clearly establish what is known, what limitations exist in the current understanding of the problem, and precisely how the research endeavor makes a significant and unique contribution to the existing body of methodological knowledge.

3. Research Hypotheses

The review of prior work logically precedes the presentation of research hypotheses that are specific to the proposed simulation study. These hypotheses should serve precisely the same role as would be expected in any novel research study in another substantive discipline. That is, the theory should be clearly articulated; the logical sequence of steps in deriving and articulating the research hypotheses should be delineated; and the hypotheses should be stated in a way that are both amenable to empirical evaluation and facilitate inferences that can be drawn back to theory. The research hypotheses commonly serve to link the theoretical developments and prior empirical findings with the justification for the experimental design being presented. Any disjuncture among theory, hypotheses, and design can undermine the validity of the entire set of obtained findings.

4. Justification for Using Simulation

When comparing alternative methods to answer questions about properties of quantitative techniques, analytical derivations and computer simulation studies each have their associated merits and shortcomings. A main drawback of simulation studies lies in that results can often not be generalized beyond the conditions included in the study. Conversely, analytical derivations can often be applied generally and are therefore more broad and informative. The limitations of Monte Carlo simulations are aptly summarized in a 1979 publication policy from the Psychometric Society, which states, “Monte Carlo studies should be employed only if the information cannot reasonably be obtained in other ways” (Psychometric Society, 1979, p. 133).

However, at the same time simulation studies are not inherently flawed; rather, their limitations may merely necessitate further justification for their use. Specifically, the researcher should address why the question under investigation could not be assessed using more formal and general methods of analysis. Common reasons motivating Monte Carlo simulations include a comparison of properties of multiple methods, perhaps under non-ideal scenarios (e.g., non-normality, small samples); sensitivity and robustness to violations of assumptions; or investigations of models that are too computationally complex to solve analytically (e.g., Bayesian estimation or methods estimated by numerical integration). Simulations may also be justified for illustrative or didactic purposes, as results can be summarized to show asymptotic or frequentist properties that may be difficult to grasp through mathematical representations. In short, the key point here is this: computer simulations should never be used as a substitute for thinking.

5. Selection of Manipulated Conditions and Supporting Rationale

Boomsma (2013) noted that Monte Carlo simulations are randomized experiments in which the researchers possess control over the manipulated conditions. Researchers have full ability to choose both the conditions varied in the simulation as well as the levels of those conditions. Some examples

of commonly manipulated conditions include sample size, the distribution of the outcome variable (or their residuals), model size, severity of misspecification, and type of estimator. When conducting a Monte Carlo simulation, all experimental design conditions should be clearly reported to aid in assessing the generalizability of the findings. For example, sample size might be a manipulated condition in the study and levels of this condition could be 100, 250, and 1000. Conditions that are relevant but that are not manipulated should also be noted (e.g., model size is kept constant and each model has five predictors). Importantly, a rationale for the selected conditions should also be provided. If sample sizes of 100, 250, and 1000 were included in the study, why are these values important to consider? Are these values that are commonly seen in this type of model or in a particular area of application in which the model is commonly used? In a simulation study intended to inform practice, choosing values that map directly onto common sample sizes seen in relevant literature can add to the contribution of the proposed simulation study. For that reason, it is unwise to select levels that correspond to more superficial characteristics, such as roundness or even spacing, as the extent to which these conditions generalize to real-life data analytic situations can come into question. A good source for such supporting evidence for levels can be found in review papers, empirical studies, and meta-analyses.

It is particularly important for researchers to closely consider the eventual external validity of their study. In theory, it might appear that manipulating many conditions with each having many levels enhances a simulation. In reality, this can sometimes have the opposite effect, making the results of a study difficult to interpret and the reporting of results (covered in Desideratum 11) quite difficult. As a simple example, if there are six independent variables under study, each of which is defined by four levels, there are 4096 unique cells in the design. Researchers should strive for parsimony when crafting the experimental design of a simulation study. That is, relevant conditions should be manipulated, but the number of conditions and constituent levels should remain targeted. This parsimony will allow for specific recommendations without an overwhelming number of tables and graphics.

6. Defining Dependent Variables

The dependent variables in a Monte Carlo simulation assess the statistical properties of the fitted models under the manipulated conditions. Dependent variables can assess various aspects of the performance of fitted models, whether at the level of the model, a specific parameter, or a given case. For example, a researcher might include model-specific information (e.g., R^2 , data-model fit indices like RMSEA or CFI, log-likelihood values, entropy), parameter-specific information (e.g., relative or raw bias, root mean squared error, Type-I error rate, empirical power, confidence interval coverage), or case-specific information (e.g., DFBeta, DFFit, individual contribution to the log-likelihood). The level of specificity (model, parameter, or case) of the dependent variable should correspond to the motivating research questions (as discussed in Desiderata 1 through 3). The definition of the quantities under investigation should be also be described, as computational formulas and associated terminology are not always consistent across software or disciplines.

7. Data Generation Model

Because the data in Monte Carlo simulations are generated according to the specifications set forth by the researcher, it is imperative to describe the data-generating model clearly. Often in simulation studies, the model used to generate the data is considered as *truth*, and the performance and statistical properties of the fitted models is often determined by how closely the estimates correspond to the model's chosen population parameter values. For this reason, researchers should report the type

of model used to generate the data either graphically (e.g., path diagrams), with equations (e.g., with regression or multilevel models), or both. As suggested by Paxton, Curran, Bollen, Kirby, and Chen (2001), the model should be reflective of the types of models seen in empirical journals so that the results are as realistic and generalizable as possible. The population values of the parameters should be disclosed, and the rationale behind those values should be justified; this justification might relate the population values to prior empirical findings, such that the generated data mimic data seen in real-life.

Researchers should also be sure to generate data from a model that could realistically be fit to the data. For example, if a researcher conducting a simulation is interested in small sample performance of particular estimators in the structural equation modeling framework, the model should be reasonably small and should not contain complex features such as latent variable interactions or categorical latent variables (e.g., as in mixture models). If 40 observations are generated from such a model, it would be unlikely that any fitted model would converge, as the complexity of the data-generating model does not match the interest of the study (this guideline would, of course, not apply if the primary interest of the study was in investigating which estimator converges most often). A similar argument exists for selecting population values for parameters. As stated previously, for the results to be as generalizable as possible, it is important that parameter values are reasonable and within the bounds of values that are observed in empirical research. Transparent descriptions and rationales for these selections can help readers evaluate the utility and generalizability of the results to how the method(s) of interest is applied in empirical contexts.

8. Data Generation Procedures

Related to Desideratum 7, because the data in Monte Carlo simulation studies are generated by the researcher, it is vital that the procedures used to generate the data be fully transparent to readers and to other researchers who might wish to replicate the findings or extend the simulation study to alternative manipulated conditions. It is important to note the type of distribution from which data were generated and the arguments of the distribution. For example, variables' data could be generated from a standard normal distribution or a normal distribution with a mean of five and a standard deviation of ten. This type of description can be especially important with discrete or non-normal outcomes. Binary variables can be generated directly by generating random variables from a Bernoulli or binomial distribution; alternatively, they can be obtained by generating data from, say, a continuous normal distribution and discretizing the values. While both are effective for generating binary variables, the underlying assumptions may not be equal (e.g., polychoric correlations assume an underlying latent normal distribution which may only apply to one of these methods) and could unintentionally affect results. As another example, when researching missing data mechanisms, a well-articulated description of the data generation process can help readers fully understand how the appropriate missingness mechanism (e.g., missing at random) was implemented. Similarly, if studying non-normality, it is important to know how non-normality was induced so that readers can assess whether the type of generated non-normality occurs frequently in particular areas of application.

Details about the specific software used should also be reported. Researchers should report the software used to generate the data, the software version number, any pre-programmed procedures that assisted in the generation of the data, and the seed number or seed assignment algorithm used in the generation of the random variables. Listing each of these facets ensures that the data generation is both transparent and reproducible. It also can help explain differences if readers attempt to replicate the results but use alternative programs or procedures.

The recent migration of publication to online formats has also allowed for supplemental materials to be included with studies more readily. We strongly encourage researchers to provide the data

generation code as an appendix, online supplemental material, or on a personal website linked in the published manuscript. This provision helps other researchers to fully understand how the study was conducted, and to be able to replicate and extend the study to additional contexts so that the phenomena of interest can be more comprehensively understood.

9. Description of Fitted Models

The overarching goal of Monte Carlo simulation studies can vary widely. In some studies, researchers are interested in studying the statistical properties of estimators or models under sub-optimal conditions (e.g., small samples sizes) to determine if parameter values can be recovered when the model is correctly specified. In other cases, researchers might knowingly fit models that are misspecified in some way (e.g., incorrect relations between variables, omitted variables, violated assumptions) to generated data to investigate statistical properties of model, estimators, or assumptions under such conditions. Because of these two different goals (which are not mutually exclusive), the models being fitted to the generated data must be fully described.

If knowingly misspecified models are fitted to the generated data, the type and magnitude of the misspecification should be elucidated. For example, if studying the effect of omitted variables from a particular model, it should be clear which variable(s) present in the data generation model have been omitted in the fitted model, the location of these variables in the model, and the anticipated size of the impact this omission will have. If data are generated from non-normal distributions but a particular estimator assumes (multivariate) normality, this should similarly be made clear. If there are conditions in which the fitted models are identical to the data-generating model, this also should be clearly stated as the statistical properties would be expected to be well-behaved in such circumstances.

Related to Desideratum 8, the software procedures for fitting such models should be included. This includes the name of the software, the procedure name, any non-default options, and the software version number. If multiple software programs are used because, for instance, certain estimators or options are not available in particular programs, this should also be mentioned so that readers are aware of the limitations present in various software options for the model type(s) of interest.

10. Analysis of Dependent Variables

A clear interpretation of the dependent variables should be provided, as should the standards for their evaluation. For example, if relative bias is an outcome of interest in a study, researchers should mention which values are considered to be indicative of poor behavior (e.g., relative bias larger than $|10\%|$ is problematic). Other outcomes may be assessed in multiple ways. As an example, Type-I error rate can be calculated directly by including a variable with null relations in the model and recording the number of replications in which the parameter estimate's test leads to a non-null inference at some pre-specified α level. Alternatively, some studies might prefer the confidence interval coverage metric, where the number/proportion of replications in which the population value is contained within the confidence interval (or pre-specified width) is tracked.

The method of analyzing the dependent variables should also be discussed. Considering that Monte Carlo simulations are experiments with manipulated independent variables, some researchers opt to fit analysis of variance (ANOVA) models to their simulation design in order to more precisely pinpoint which conditions affect the dependent variable(s) under investigation. This type of approach is popular because it also allows testing of interaction effects, which can be difficult to see visually or in tabular form. If ANOVAs are conducted, researchers must specify whether they focus on statistical significance (which may not be informative if a large number of replications is

used) or effect size measures. If the latter, then the criteria used for judging small, medium, and large effects should be listed.

Details about data analytic problems in the simulation are also important to provide. Specifically, this could include aspects like how non-convergent replications or inadmissible solutions (e.g., Heywood cases in structural equation modeling) were handled (e.g., removed from the study, with additional replications run), how potentially outlying replications were dealt with (e.g., whether to include or re-run a replication with 5000% bias or to summarize the condition with the median instead), and the unit of analysis used for analyzing the dependent variable (replications within each cell or cell means pooled across replications).

11. Presentation of Results

When a Monte Carlo simulation is the approach taken in a quantitative study, the number of manipulated conditions and the number of dependent variables typically produces a sizeable amount of results. If all of these results are exhaustively tabled, the overall conclusions can often be difficult for the reader to ascertain and the ultimate take-home message can be lost. Instead, researchers should strive to present their results as efficiently as possible. Dependent variables that did not display much variation across manipulated conditions can be described as such with a sentence or two in text and do not usually call for accompanying tables or plots. A concise sentence may also suffice for results that have been shown in previous simulation studies or that could be discerned by previous analytical work. Tables and figures should be reserved only for the most interesting or novel results. Plots are generally more advisable when the researcher wants to express broad or general trends in the results; they can also be particularly useful for viewing an interaction effect among simulation conditions. Showing that the relative bias of estimates from different estimators decreases as sample size increases is one such example. This type of information is quickly seen in graphical form whereas it might take additional processing by the reader if presented in a table. Tables are best used if the researcher wants to draw attention to specific values rather than broad trends, especially if there are relatively few conditions. For instance, if one is studying the performance of structural equation model fit indices under various types of non-normality, the specific values may be more salient to present as a function of which conditions the fitted models indicate good or poor data-model fit, particularly in relation to other published guidelines.

Another important consideration when summarizing results of a Monte Carlo simulation study concerns how to aggregate the manipulated conditions. Monte Carlo studies can manipulate half a dozen or more independent variables, and reporting across each separate independent variable may make the results difficult to synthesize. Therefore, in such cases it might be helpful to aggregate the results over some of the conditions that yielded less interesting results. For example, if a multilevel model simulation manipulated the intraclass correlation, level 1 sample size, and level 2 sample size, and was concerned with the bias of the variance component estimates, the results for level 1 sample size conditions might not yield very different results across different conditions. Hence, instead of reporting the level 2 sample size results separately by level 1 sample size conditions, it is likely more parsimonious to aggregate over level 1 sample size when describing the main effect of interest. Related to Desideratum 8, less insightful and/or more detailed results can be placed in an appendix or on an author's personal webpage.

12. Summary of Key Findings and Recommendations

The ultimate goal of a Monte Carlo simulation is to advance the literature regarding the appropriateness, robustness, and/or utility of statistical methods under certain conditions. Once the

relevant results are summarized and reported, researchers should not allow readers to interpret the results for themselves, but rather should make explicit recommendations and suggestions based on the findings. Although simulation studies will necessarily be limited in their ability to generalize the results beyond the conditions included in the study, the researchers conducting the study have the most knowledge about the simulation that was conducted, including where problems were encountered. If the study compares various estimators under certain conditions, the researcher should provide readers with guidelines that are as clear as possible regarding when each estimator should be used. If the study addresses issues of model misspecification, a discussion of the robustness of the method(s) should be given so that readers know when it is and is not appropriate to apply the method(s). By the end of the paper, readers should feel like they have a satisfying answer to the research questions outlined in Desideratum 3, or at least an understanding of why a clear answer could not be given at the end of the current investigation and what steps might be taken to obtain clarity in the future.

References

- Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling*, 20, 518–540.
- Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, 8, 287–312.
- Psychometric Society. (1979). Publication policy regarding Monte Carlo studies. *Psychometrika*, 44, 133–134.

21

Multidimensional Scaling

Cody S. Ding and Se-Kang Kim

Multidimensional scaling (MDS) is a multivariate statistical method for estimating the scale values along one or more continuous dimensions such that those dimensions account for distance measures defined over pairs of variables or items (also called objects or stimuli). It has been used to study such things as dimensions underlying perceptions of human speech, patterns of vocational/academic interests, and growth over time in reading and math achievement. It can also be used as a data visualization method to depict data structure in a two- or three-dimensional space. We will limit ourselves to discussion of analyses based on Euclidean distance models, which is based on distance measures or coefficients. While the distance measures can be defined in different ways according to different purposes, the distance measure most commonly used is the one that is computed over pairs of variables using a rating scale such as a Likert-type scale. MDS can be used for different types of studies. Perhaps the most typical form of MDS analysis that is traditionally used and most familiar to researchers is akin to factor analysis, in which the purpose is to identify the attributes or factors (dimensions) along which variables are perceived to vary and that account for the data (a perception application). However, MDS can also be used for profile analysis in cross-sectional studies, where the behaviors are measured at a single time point (e.g., score on a vocational interest scale.) and the purpose of the analysis based on the distance measure is to identify dimensions that point toward one or more within-person patterns needed to account for the associations among the variables. In longitudinal studies, which are a relatively recent extension of cross-sectional MDS methods, the goal is to find patterns of growth, decay, or change that account for the associations over time or the occasions based on a distance measure. Thus, the data consist of a single variable (e.g., math achievement) measured at several time points.

Although most applications of MDS are exploratory, designed to uncover patterns along two or more dimensions accounting for the data, they can also be used to test *a priori* hypotheses about the dimensional or spatial structure that accounts for the data. There are constrained versions of MDS designed for fitting *a priori* hypotheses, and the fit measure used to evaluate the dimensional or spatial structural hypotheses is typically the correlation coefficient. Cross-sectional and longitudinal MDS analysis can serve to generate hypotheses that will later be assessed for confirmation through methods discussed in other chapters (e.g., structural equation modeling, hierarchical linear modeling).

Table 21.1 Desiderata for Multidimensional Scaling.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
<i>General</i>	
1. Describe the competing theories and prior research results leading to the study, including predictions about the dimensions or spatial configuration of variables needed to account for the data.	I
2. State the purpose to which MDS will be used; for example, investigate perceptual dimensions of variables, visualize data structure in a two- or three-dimensional space, recover within-person patterns in cross-sectional data, or explore patterns of growth and change in longitudinal data.	I, M
3. Describe the sample of participants; describe and justify the population to which results may be generalized.	M
4. Describe the distance model of the data and the function of dimensions.	M
5. Describe the measures of fit that will be used to compare models or to decide on the dimensionality of the final solution.	M, R
6. Describe and, if possible, justify the rotation of the reported solution.	M, R
7. Describe the variables used and to which results can be generalized.	M
8. Describe the distance measure used.	M
9. Explain how missing data, if any, were handled.	M
10. Justify the final model selected, including the dimensionality of that model.	R
11. Report the scale values of the final solution in table and/or graphical form. Explain and justify the interpretation of the dimensions as their scale value patterns or the spatial configuration of variables.	R
<i>Within-Person Patterns in Cross-Sectional Studies</i>	
12. Describe and justify the variables under study.	M
13. Explain how missing data, if any, were handled.	M
14. Describe distance measure (measure of association between pairs of variables) selected for the study. Report descriptive statistics on the variables and report the distance matrix if needed.	R
15. State and justify the distance model used, the method of estimating the parameters in the model for distance measure, and the fit measure(s).	M
16. Report the scale value of the final solution in tables or graphs. Explain and justify the interpretation of the dimensions as their scale value patterns or the spatial configuration of variables on which the final conclusions are based.	R
17. Justify MDS over alternative analyses. Report or discuss parallel results from methods related to MDS (e.g., Q-factor analysis, cluster analysis).	M, R
<i>Growth Patterns in Longitudinal Studies</i>	
18. Describe and justify the proximity measure (distance or correlation between pairs of time points) selected for the study.	M
19. Describe the time points and model under study and describe how the scale value patterns are interpreted with respect to growth or change.	
20. Explain how missing data, if any, were handled.	M
21. Explain and justify the interpretation of the dimensions as their scale value patterns or the spatial configuration of variables with respect to growth or change on which the final conclusions are based.	R, D

* I = Introduction, M = Methods, R = Results, D = Discussion.

Kim, Frisby, and Davison (2004) described the application of MDS to cross-sectional data; Ding, Davison, and Petersen (2005) described the application to longitudinal data; and Kim, Davison, and Frisby (2007) discussed the translation of hypotheses generated by MDS into structural equation models. More thorough treatments of MDS models can be found in Borg and Groenen (2005), Cox and Cox (2001), and Davison (1983). SAS and SPSS contain MDS programs (ALSCAL, Takane, Young, & de Leeuw, 1977; PROXSCAL, Commandeur & Heiser, 1993). Other programs include SMACOF, which is embedded in R (de Leeuw & Mair, 2009), MULTISCALE (Ramsay, 1977), SMALLEST SPACE ANALYSIS (Lingoes, 1989), and PROSCAL (MacKay & Zinnes, 2005).

In describing key methodological issues in MDS studies, we begin the table of desiderata and subsequent elaborations by describing those elements that are common to MDS analyses (i.e., data structure or pattern in perception studies, visualization of data structure, cross-sectional, and longitudinal studies), followed by a discussion of issues specific to cross-sectional and longitudinal applications.

1. Theory and Prior Research

While MDS is usually exploratory, a solid grounding of the study in theory and prior research is no less necessary. Existing theory is often insufficiently precise for purposes of specifying hypotheses to the degree of precision required by confirmatory analyses—hence the need for exploration.

The theory and prior research in the specific content area leading to the current study need to be explained. Competing theories, if any, should be included in the explanation. If MDS analysis is used in the previous research or the theory relates to issues of dimensionality, the explanation should include any dimensions or any spatial configuration suggested by that prior literature. The prior literature might be used to guide the number of dimensions included and the substantive interpretation of that final solution. In some cases, confirmatory MDS analysis can be conducted given this prior theory or research. It might also guide the selection of a respondent population or a variable population to be sampled. Further, it might suggest additional data to be collected for the purpose of confirming interpretations of the dimensions or the spatial configuration. In fact, this part of the write-up is no more different from the write-up using other methods such as regression or structural equation modeling.

2. Purpose

In their write-up, authors need to explain the intended purpose for utilizing MDS. The stated purpose will differ depending on whether MDS is being used to study perceptions of stimuli, underlying data structure, variables collected in a cross-sectional study, occasions in a longitudinal study, or some other entities. In perception studies, the researcher often intends to recover dimensions accounting for the perceptual judgments as well as a description of individual differences in the use of those dimensions. However, if MDS is used as a data map for examining data structure in a two- or three-dimensional space, the focus should be on examining cluster patterns of variables. In this case, describing dimensions may not be necessary. In cross-sectional studies, the researcher is often interested in describing dimensions that account for within-person variation, in which case, the researcher should so indicate. If the cross-sectional data are being collected for some other purpose, that other purpose should be described. Likewise, longitudinal data are often collected for the purpose of uncovering patterns of growth, decay, or change, and if so, that intent should be stated. If the data are being collected for purposes of hypothesis testing, any planned confirmatory analysis needs to be described.

3. Sampling of Participants

The Methods section must include a description of the population sampled and the sampling method used, like the studies using any other methods. This will determine the limits of generalizability with respect to participants. The description should include a discussion of important subsamples that might vary in their perceptions of stimuli, their data structure, their patterns of scores on cross-sectional variables, or their growth trajectories in longitudinal studies.

4. Describe the Distance Model and the Function of Dimensions

The researcher needs to describe the model, or models if several were tried, for the data. This description should indicate whether the model is nonmetric or metric. Nonmetric scaling assumes that the data form an ordinal scale. On the other hand, metric scaling assumes that the data constitute either an interval or ratio scale. Individual differences parameters, if any, should be described. The description should indicate whether the model assumes that the data are related to Euclidean or non-Euclidean distances, and if Euclidean, whether Euclidean or squared Euclidean distances. If models with varying numbers of dimensions were fitted to the data, the range of dimensionality should be reported.

Consider the two models below for the dissimilarity δ between variables:

$$\delta_{ii'} \approx f \left[\sqrt{\sum_k (x_{ik} - x_{i'k})^2} \right] \quad (1)$$

$$\delta_{ii'p} \approx f \left[\sqrt{\sum_k w_{pk} (x_{ik} - x_{i'k})^2} \right] \quad (2)$$

The first is a nonmetric model that assumes that the dissimilarity for variables i and i' is a monotonically increasing function f of Euclidian distances between points for variables i and i' whose locations in a K -dimensional space are given by the coordinates x_{ik} and $x_{i'k}$, respectively. The second assumes that the dissimilarity data for person p is a weighted Euclidean distance function with individual differences weights w_{pk} representing individual differences in the saliences of the perceived dimensions. Inclusion of the monotone function in Equation (1) implies weaker, ordinal assumptions about the dissimilarity data, but the model does not accommodate individual differences in the stimulus perceptual process. The model of Equation (2), however, makes stronger, ratio scale assumptions about the data but allows for individual differences in the perceptual process through inclusion of the weight parameter w_{pk} . Models vary in several respects such as their assumptions about the measurement scale of the dissimilarity data $\delta_{ii'}$ and their assumptions about individual differences.

5. Fit and Dimensionality

The various fit measures used to evaluate and compare solutions should be described. Criteria used to evaluate those fit measures should also be described. For instance, Kruskal (1964a, 1964b) provided guidelines for the least squares fit measure STRESS. A distinctive "C" or "U" shape of the two-dimensional configuration can be an indication that only one dimension is really needed to account for the data. It should be noted that STRESS values decrease as the dimensionality increases, and thus, it is also important to consider the portion of variation in data account for by the dimensionality (e.g., R^2) and the meaningful interpretation. In confirmatory MDS analysis, AIC or BIC can be used to compare different models for determining the best fitting model.

Although it has been suggested that at least five stimuli are needed per dimension in order to achieve satisfactory precision, some seemingly meaningful solutions have been obtained with as few as three stimuli per dimension in cross-sectional and longitudinal applications. In addition, if the purpose of the analysis is to visualize data structure in a two-dimensional space, this guideline may not be necessary. Some computer programs print a warning if there are few stimuli per dimension and will not compute a solution if there are too few.

Like most iterative algorithms, MDS algorithms may fail to reach an optimal solution for one of several reasons, including local minima, solution degeneracy, or an inadequate number of iterations. Researchers should familiarize themselves with these problems, methods for detecting such problems, and methods for avoiding them. Obviously, non-optimal solutions should not be reported and steps taken to avoid such problems should be described. One rule of thumb to prevent the problems is to have enough iterations (e.g., 200 or more) for scaling.

6. Rotation of the Solution

Except in certain special cases, MDS solutions are subject to the same rotational indeterminacy as are exploratory factor solutions. Unfortunately, there are no widely accepted algorithms for optimizing the interpretability of the solution through rotation as in factor analysis. When interpreting the spatial configuration of stimuli, the rotation may not matter. For instance, if the variables form a clear clustered configuration in two dimensions, that configuration is invariant with respect to rotation and, consequently, rotational indeterminacy poses no limitation. On the other hand, when interpreting dimensions, the dimensional interpretation applies only to a particular rotation.

In some MDS models, most notably those based on the weighted Euclidean model, the rotation is determinate except in certain special cases. When the solution is based on a model in which rotation cannot be performed without loss of fit, this rotational determinacy should be noted.

When interpreting dimensions of the solution, rather than the more general spatial configuration of the stimuli, the researcher needs to justify the chosen rotation or acknowledge its indeterminacy as a limitation. Alternative dimensional interpretations corresponding to rotations of the solution need to be recognized. In perceptual studies, the researcher may wish to employ an analysis based on an individual differences model in which the rotation is generally determined because the dimensions are tied to the individual weights that are fixed.

7. Sampling of Stimuli or Variables

The Methods section must describe variables included in the study. MDS dimensions are those along which stimuli or variables vary. In a study of occupational perceptions, for instance, the resulting dimensions are likely to include a dimension of occupational safety only if the sample of occupations includes both safe and dangerous occupations. The description should indicate whether the sample of variables is considered *fixed* (i.e., all stimuli of interest) or *random* (i.e., only a sample of stimuli of interest). Because the selection of stimuli can seriously influence the interpretation of dimensions, possible effects of stimulus selection may need consideration in the Discussion section.

8. Describe the Distance Metric

In the most traditional form of MDS analysis, the input data are based on dissimilarity measures for each pair of stimuli that can come from a variety of sources. For instance, the participant might be shown two statements/stimuli and asked to rate them on the following 7-point scale ranging from *highly similar* to *highly dissimilar*. Rating scales, however, are not the only possible task. For instance,

the participant can be shown three stimuli/statements and then, from among the three stimuli, be asked to select the two that are most alike and the two that are least alike. Measures from indirect judgment tasks can also be used. For instance, the researcher can briefly present pairs of statements and then ask the participant to indicate whether they were the same or different. The number of times two statements are confused, that is, incorrectly identified as being the same, can be considered a measure of their similarity. Regardless of how the dissimilarity measure is derived, the researcher must describe the task and how responses were scored to obtain a dissimilarity measure for an MDS analysis. This description should include the directions given to respondents or, at least, a summary of those directions. If judgment tasks are used to obtain the dissimilarity measure, the order of stimulus presentation, both within and across pairs, can potentially have an effect on judgments and any steps taken to control order effects should be described. If the MDS analysis algorithm assumes that, within the limits of random error, similarity is symmetric (i.e., the similarity of pair (A, B) is the same as the similarity of (B, A) and that the order of presentation does not matter), the researcher would need to describe any asymmetries in the data and the method of handling such asymmetries.

9. Missing Data

As the number of stimuli/statements increases, the number of stimulus pairs increases rapidly. If n is the number of stimuli, the number of stimulus pairs is $n(n - 1)/2$. If the number of pairs is large, an incomplete data collection design can be employed to reduce the number of judgments by any one participant. For any one participant then, some judgments will be missing by design. Such missingness needs to be described, and its potential impact on the results should be discussed. Data may also be missing not by design. This missingness should also be described along with any steps to handle the missingness such as pairwise deletion or missing value imputation.

10. Final Model

Often, the analysis will include a comparison of several models. Almost all research includes a comparison of models of varying dimensionalities. It may include models with and without constraints on scale values (e.g., in confirmatory MDS), models that do and do not include individual differences parameters, or models that are metric with those that are nonmetric. Generally, the models are compared in terms of parsimony, fit to the data, interpretability of the dimensions, and replicability of dimensions across samples. All other attributes being equal, a model is preferred if it contains fewer dimensions or freely estimated parameters (parsimony), better fit to the data (fit), dimensions all of which are substantively interpretable or an interpretable spatial configuration of stimuli (interpretability), and dimensions or configurations that appear in the solutions of several samples (replicability).

11. Final Solution

The scale values of the final solution should be reported in tabular form, graphical form, or both. Since the MDS solution is interpreted in terms of the spatial configuration, graphical presentation is essential. Reported results should also include one or more fit measures. If an individual differences model has been employed, estimated individual differences parameters should be reported in graphical or tabular form. If the sample size is large, however, these individual differences results may be reported in summary form (e.g., means and standard deviations of estimated individual differences parameters).

Following the presentation of the solution itself, the Results section should include results of any analyses that aid in the interpretation of the solution. Often these include various correlational analyses.

For instance, in a hypothetical study of job perceptions, the scale values along Dimension 1 might be correlated with the median salary associated with each job in an attempt to determine whether Dimension 1 can reasonably be interpreted as reflecting salary. A cluster analysis of estimated dimension scale values might reveal distinct groupings of occupations. While the interpretation of the solution is ultimately subjective, it can be aided by additional data and analyses.

12. Variables in Cross-Sectional Studies

Although MDS has traditionally been used to study stimulus perceptions or data structure as described above, it can also be used to study patterns of scores (i.e., profile) in cross-sectional data. When MDS is used for profile analysis, the scale value pattern in a given dimension is interpreted as a core profile that encapsulates all person responses on a set of variables included in a sample. A pattern is defined as peaks and valleys appearing over a set of variables. The peaks usually depict high scores on the measuring variables, whereas the valleys for low scores on them. Changing the variable set can alter the score patterns in the data. Therefore, the Methods section must describe the variables included in the set of measures analyzed and explain the rationale for their selection. Often the variables are scales which constitute an assessment inventory or test battery. The scales are commonly administered together and are often reported as a profile of scores in clinical, counseling, educational, or industrial/organizational psychology (e.g., the scales in an interest inventory or the scales in an intelligence test battery).

In some cases, the possible effects of changing the composition of the variable set will need to be discussed. Where the composition of the variable set is an issue, the researcher may want to empirically study the effect of adding or deleting variables. For instance, a researcher might examine the stability of Big Five personality scale patterns across two personality inventories, one of which included only Big Five personality scales and one of which included Big Five scales embedded in a larger set. If the variables have been sampled from a larger domain, the researcher will need to describe the larger domain from which the variables originate and discuss the generalizability of patterns to variable sets in the larger domain.

13. Missing Data in Cross-Sectional Studies

For any given person, data on some variables might be missing. The MDS distance measures can often be computed within existing computer packages (e.g., SAS, SPSS), and these packages often include listwise and pairwise options for handling missing data. Such options can be justified if the data are missing completely at random (MCAR). When the MCAR assumption is satisfied, pairwise deletion uses larger samples for computing the results and therefore yields sample estimates of distance measures with smaller standard errors. When the sample size is large, however, listwise deletion should yield distance measures with sufficiently small standard errors and will ensure that every distance measure is computed on the same sample of data.

If the MCAR assumption is not defensible, however, the researcher may want to employ some form of data imputation (usually multiple imputation), before the proximity measures are computed. Existing statistical packages typically offer several imputation options. The option chosen should be described and justified.

14. Describe the Distance Measure in Cross-Sectional Studies

Even casual inspection of the distance module in any of the common statistical packages will reveal that there are a large number of statistical measures of association that might be employed.

The correlation and covariance statistics are probably the two most widely known in the social, behavioral, and education sciences. The choice of distance measure must be justified. In our opinion, the strongest basis for justification begins with a model of the raw data from which a distance measure can be derived. An example of such a model is given below.

In MDS, one plausible distance measure is the squared Euclidean distance which can be computed from the raw data for variable pair (v, v') as follows:

$$\delta_{vv'}^2 = \frac{\sum_p (y_{pv} - y_{pv'})^2}{P} \quad (3)$$

In words, the squared Euclidean distance measure of distance for variables v and v' is the squared difference between the score of person p ($p = 1, \dots, P$) on variable v and v' averaged across all persons (however, SPSS computes the sum over all persons, not the average). Note that P represents the total sample size.

The most well-known measure of association, the Pearson product moment correlation coefficient, is closely related to the squared Euclidean distance. If the variables are in z -score form (i.e., variables z_{pv} and $z_{pv'}$), then the squared Euclidean distance proximity measure has the following form:

$$\delta_{vv'}^2 = 2 - 2r_{vv'} \quad (4)$$

Equation (4) says that, when the squared Euclidean distance proximity measure is computed from standardized variables, the squared Euclidean distance proximity measure is linearly but inversely related to the correlation coefficient. In MDS, this means that an MDS of correlations among variables will yield exactly the same solution as an analysis of squared Euclidean distances computed from variables in standardized form if two conditions hold: (1) the correlations are treated as similarities whereas the squared Euclidean distances are treated as dissimilarities and (2) both proximity measures are treated as ordinal or both are treated as interval level data points.

In our opinion, unless there is very good reason to do otherwise, researchers should base analyses of cross-sectional data on squared Euclidean distances or correlation coefficients, at least if the researcher intends to analyze a variable-by-variable matrix rather than a person-by-person matrix. Either of these measures can be justified from a plausible, explicit model of the original variables, Y_v . That model is described in the next section. An explicit model of the original variables can not only serve as the basis for justification of a proximity (distance or correlation) measure, but also it can enrich the interpretation of the resulting MDS dimensions. A table of proximity measures should be reported in the results section to facilitate later re-analysis of the data and meta-analysis.

15. State Model of the Observed Variable in Cross-Sectional Studies

In cross-sectional applications of MDS, the phrase “model of the data” can mean one of two things: a model for the original variables Y_{pv} or a model of the distance measure. In this section, we are primarily concerned about a model of the original variables Y_{pv} , but in some cases the model of the original variables can be used to derive a model for the distance measure. If the researcher has a model for the original data, the model should be stated. The model constitutes a statement of the assumptions on which the analysis is based. If there is no such model, and frequently there has not been one in cross-sectional applications, then the conditions under which the analysis is appropriate are unstated, seemingly unknown, and impossible to evaluate. Without such a model, there is no formal connection between the resulting MDS scale values and the original data thus precluding formal explanations of the original variables in terms of the MDS solution.

One possible model from which a distance measure can be derived is the Profile Analysis via Multidimensional Scaling (PAMS) model:

$$y_{pv} = c_p + \sum_k w_{pk} x_{vk} + e_{pv} \quad (5)$$

where y_{pv} is the observed score of person p ($p = 1, \dots, P$) on variable v ($v = 1, \dots, V$), which represents the element in row p and column v of the data matrix; c_p is a level parameter which indexes the overall height of person p 's profile, $c_p = \frac{\sum_v y_{pv}}{V}$, which is in fact unweighted average for person p ; the scale value x_{vk} along dimension k constitute a row vector \mathbf{x}_k of contrast coefficients that depict a pattern of scores in given dimension k ; w_{pk} is a weight for person p on dimension k that indexes the degree of match between the pattern of person p 's observed scores and the pattern of the scale values in vector \mathbf{x}_k ; and e_{pv} is an error term.

In essence, the PAMS model in Equation (5) represents each person's row vector of data \mathbf{y}_p as a linear combination of the patterns \mathbf{x}_k represented as vectors of MDS scale values which construct a core profile given dimension k :

$$\mathbf{y}_p = c_p \mathbf{1} + \sum_k w_{pk} \mathbf{x}_k + \mathbf{e}_p \quad (6)$$

where $\mathbf{1}$ is a row vector of 1s and \mathbf{e}_p is a vector of error terms for person p . Readers familiar with factor analysis will recognize this as a linear model similar to that in factor analysis except that it includes an intercept term. In words, Equation (6) states that each person's row of data is a linear combination of pattern vectors \mathbf{x}_k . While multidimensional models including a person-specific intercept have long existed in the scaling literature, factor models with a random coefficient intercept are a more recent development.

Given appropriate assumptions (Kim et al., 2007) and if computed from the raw data according to Equation (3), the squared Euclidean distance measure for each variable pair will have the following form:

$$\delta_{vv'}^2 = \sum_k (x_{vk} - x_{v'k})^2 + 2\sigma^2 = d_{vv'}^2 + 2\sigma^2 \quad (7)$$

where σ^2 is the variance of the errors in Equation (5). Hence, the distance measures are a squared Euclidean distance function of parameters x_{vk} in the model; an MDS of such distance measures can be used to estimate the parameters; and the scale values in the MDS will constitute estimates of those parameters. Having estimated the parameters x_{vk} through MDS, one can use the scale value estimates and regression to estimate the individual differences parameters w_{pk} and c_p .

In addition, one must decide whether to consider the proximity data as ordinal, interval, or ratio in order to select an appropriate metric or nonmetric analysis. If there is a formal model of the original data from which the proximity measure has been derived, then the derived form of the proximity measure may determine the appropriate analysis. For instance, consider the model of the proximities in Equation (7) derived from the model for the raw data in Equation (5). According to Equation (7), the proximity data are not proportional to distances and therefore should not be treated as ratio data. It does, however, suggest that the proximity data are linearly related to squared distances and therefore could be treated as interval (or ordinal) level data for purposes of any analysis, such as that of ALSCAL, for which metric analyses include those based on the assumption of proximity data linearly related to squared distances. In most MDS analyses, however, the proximities would have to be treated as ordinal because most analytic models assume the data are monotonically but nonlinearly related to distances.

Final selection of a model also means deciding on the number of dimensions to retain. As described earlier, the decision can be based on the number of data points, parsimony, dimensional interpretability, fit to the data, and dimension replicability across samples.

16. Report Final Solution in Cross-Sectional Studies

The Results section should contain a table, such as Table 21.2 showing the scale values for the final solution, preferably with estimates of standard errors for those scale values (Ding, 2005; Kim et al., 2004). Table 21.2 shows a two-dimensional solution from an analysis of squared Euclidean distances for all possible pairs of Woodcock–Johnson Psychoeducational Battery—Revised (Woodcock & Johnson, 1989) cognitive ability cluster scales in a sample of 357 respondents. In Table 21.2, scale values that are significantly different from zero are indicated by asterisks. Various plots can visually aid understanding of the dimensions.

When the analysis is based on the PAMS model (see Desideratum 15) and dimensions are interpreted in terms of score patterns, plots of dimension scale values against variables (Figure 21.1) can be used to portray the dimension patterns. The top of Figure 21.1, Dimension 1 scale values, shows a pattern marked by relative strengths in Speed of Processing (SPR) and Comprehension Knowledge (CKW), coupled with relative weaknesses in Long Term Retrieval (LTR), Auditory Processing (APR), and Visual Processing (VPR). The second dimension shows a pattern with a relative strength in Speed of Processing (SPR) coupled with a relative weakness in Short Term Memory (STM). Note that only variables with scale values significantly different from zero in Table 21.2 were used to identify relative strengths and weaknesses along dimensions. Plots of scale values against variables (e.g. Figure 21.1) may only be useful when dimensions can be interpreted as patterns of relative strength and weakness. In other situations, other graphical forms may prove more informative.

The scale values can be interpreted either in terms of each dimension separately, as in Figure 21.1, or in terms of the overall configuration. For instance, theories positing a *circumplex* structure of variables lead to the prediction that the stimuli will fall in a circular two-dimensional arrangement. In such circumstances, Dimension 1 and Dimension 2 scale values are graphed against each other in a scatter plot and the overall configuration in the resulting plot is visually examined to evaluate whether the variables fall in a circular arrangement and whether they fall along the circle in the order predicted by theory. Whether in terms of separate dimensions or the overall stimulus configuration, the MDS scale values in the final solution should to be interpreted and related to theory.

Table 21.2 Woodcock–Johnson—Revised Ability Cluster Coordinates and Standard Errors; Standard Errors Estimated from 200 Bootstrap Replicated Samples.

<i>Observed Variables</i>	<i>Dimension 1</i>	<i>Dimension 2</i>
LTR	-1.44* (.35)	.17 (.09)
STM	.01 (.14)	-1.43* (.37)
SPR	1.04* (.27)	.51* (.19)
APR	-1.10* (.27)	.06 (.09)
VPR	-1.01* (.25)	.46* (.13)
CKW	2.46* (.60)	.08 (.09)
FRE	.03 (.09)	.15 (.12)

Statistically significant scale value estimates at $\alpha = .05$ are indicated by *. LTR = Long-term Retrieval; STM = Short-term Memory; SPR = Speed of Processing; APR = Auditory Processing; VPR = Visual Processing; CKW = Comprehension-Knowledge; FRE = Fluid Reasoning.

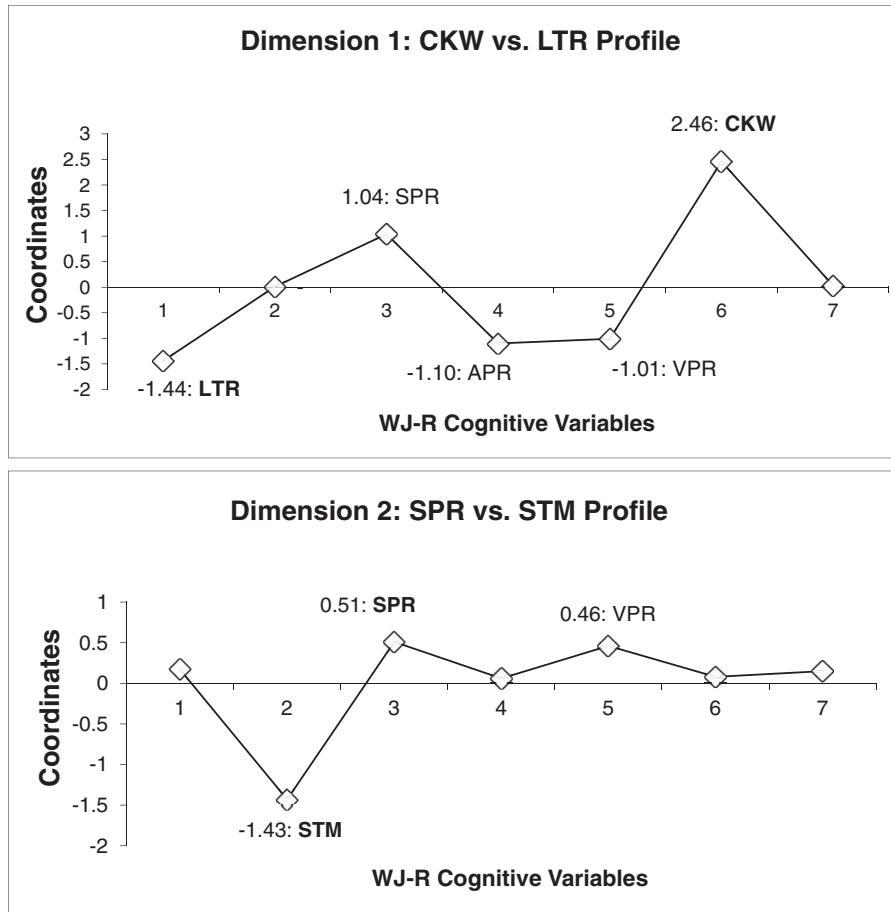


Figure 21.1 Profile Patterns of Woodcock–Johnson—Revised Ability.

LTR = Long-term Retrieval; STM = Short-term Memory; CKW = Comprehension Knowledge; VPR = Visual Processing; SPR = Speed of Processing; and APR = Auditory Processing.

17. Other Related Methods in Cross-Sectional Studies

In thinking about alternative analyses of cross-sectional data, to what analyses should MDS be compared? Because MDS and factor analysis (see Chapter 8, this volume) both yield representations of variables in a continuous space, it is rather natural to compare them. Davison (1983) described a seemingly common, if not universal, relation between unrotated components and MDS solutions in which the first (or general) component has no counterpart among the MDS dimensions but remaining unrotated components do have counterparts among the MDS dimensions. It can be argued, however, that MDS dimensions primarily describe within-person variation. This is consistent with Davison's finding that MDS dimensions contain nothing resembling a general component as the general factor primarily reflects between-persons, not within-person, variation. Because of its focus on within person variation, MDS and typical factor analysis would seem to serve somewhat different purposes.

Q-factor analysis, cluster analysis (see Chapter 4, this volume), or methods that combine the two (modal profile analysis) have often been used to describe within-person variation.

For comparison purposes, researchers may want to present alternative solutions derived with clustering or Q-factoring either to corroborate or to complement the MDS solution. Alternatively, the researcher may wish to describe why MDS was chosen over other possible analyses of within-person variation. Kim et al. (2004) contrasted cluster, Q-factor, and MDS analyses of cross-sectional data.

MDS can be used to generate hypotheses about dimensions of within-person variation, hypotheses that are subsequently tested using structural equation modeling or mixed effects modeling in a later sample (Kim et al., 2007). This approach seems especially promising in the study of variables that display patterned covariance or correlation matrices (e.g., a circumplex matrix, a simplex matrix) given that such patterning arises from factors/dimensions of within-person variation. Repeated measurements of a single variable often display a simplex structure. The analysis of such repeated measures is the topic of our next section.

18. Describe the Proximity Measure in Longitudinal Studies

Researchers need to describe their choice of proximity (distance or correlation) measure and explain why that particular proximity measure was chosen over the numerous other possibilities. As with cross-sectional data, the best justification, in our view, is one that starts with a model of the raw data, such as Equation (5), from which one can derive a proximity measure linearly or monotonically related to a distance function of the parameters to be estimated with MDS, the x_{vk} , in the case of Equation (7). Other forms of justification are presumably possible, however.

The model in Equation (5) leads to the choice of the squared Euclidean distance proximity measure defined over all possible pairs of time points and computed according to Equation (3). Under plausible assumptions, the squared Euclidean distance proximity measure computed from the raw data will be linearly related to the squared distance function of the parameters to be estimated, the x_{vk} , leading to the conclusion that MDS will provide plausible estimates of those parameters. Because the parameters have a natural interpretation in terms of change patterns, and the MDS scale values are estimates of those parameters, the MDS dimensions can be interpreted as estimates of change pattern vectors.

19. Describe Time Points and the Model in Longitudinal Studies

Whereas cross-sectional data consist of V variables measured at a single occasion, longitudinal data consist of a single variable measured at V time points. Equations (5) and (6) express the model on which the longitudinal analysis is based. When applied to longitudinal data, however, y_{pv} is the measurement of person p at time v ; c_p is an intercept for person p ; w_{pk} and e_{pv} are interpreted as before; and x_{vk} is a dimensional scale value which is the score at time v in the k th pattern of change. As the scale value pattern was interpreted as a core profile, the scale value along each dimension are interpreted as a vector describing a pattern of growth, change, or decay. Each person's longitudinal vector of scores $\mathbf{y}_p = \{y_{pv}\}$ is represented as a linear combination of K change patterns $\mathbf{x}_k = \{x_{vk}\}$. When the scale values for a given dimension are plotted against time, the plot visually displays one of the K change patterns. Figure 21.2 shows an example of distance as representation of growth rates for different time intervals

The Methods section should describe and justify the sample of time points selected for study. Typically, at least four time points or measurement occasions should be present. Because the selection of time points can seriously influence the form of the growth or change patterns, possible effects of time point selection may need to be considered in the Discussion section. Time points need not be equally spaced. For instance, in a study of gains in reading achievement, there could be

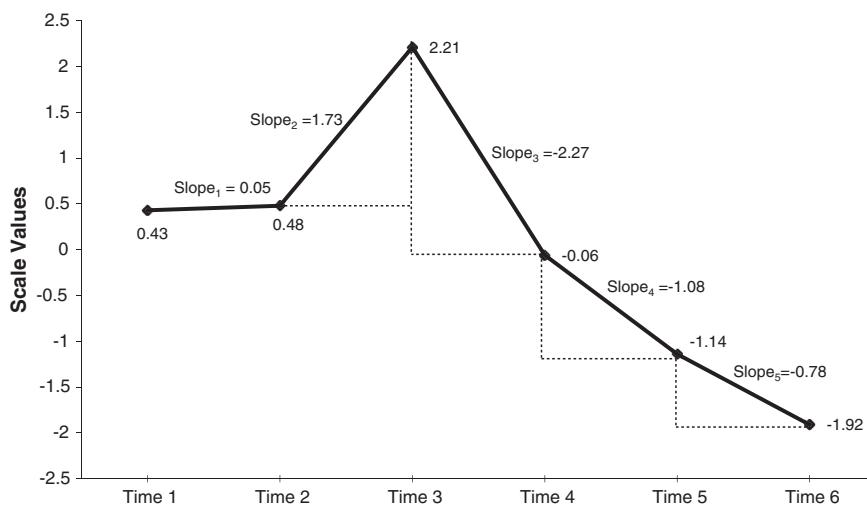


Figure 21.2 Distance as Representation of Growth Rates for Different Time Intervals.

a two-month interval between times 1 and 2, but a four month interval between adjacent time points thereafter. In any graphical representation of a change pattern plotted against time, the unequal spacing of time points should be displayed along the horizontal axis. Failing to accurately represent the unequal intervals between time points will distort the graphical representation of change rates over intervals.

Model justification can have several aspects. First, if the proximity measure is derived from a model of the raw data, then the plausibility of the raw data model needs to be considered. Second, model justification involves an explanation of the assumed form for the proximity measure. If, however, the researchers have no model of the raw data or no way of deriving a proximity measure from it, some other form of justification must be provided. In either case, the explanation must include a justification for the assumed measurement level of the proximity data. That is, does the researcher assume the proximity measures to be ordinal (i.e., monotonically related to a distance function of the scale values), interval (i.e., linearly related to a distance function of scale values), or ratio (i.e., proportional to a distance function of scale values)? This assumption will determine whether the MDS analysis will be of the nonmetric or metric form.

Finally, in explaining the model, the researcher must justify the number of dimensions retained in that MDS solution. As explained earlier, the justification may be based on the number of time points, model fit, parsimony, dimension interpretability, and dimension replicability across samples of respondents.

20. Missing Data in Longitudinal Studies

Because longitudinal data are often incomplete, the researcher must describe and justify the method of handling missing observations. Some programs offer the researcher the option of listwise or pairwise deletion in the computation of a proximity measure, such as the squared Euclidean distance. Either pairwise or listwise is readily justified when data are missing completely at random (MCAR). In our opinion, however, data are seldom missing completely at random, and the MCAR assumption is even more difficult to justify for longitudinal data than for cross-sectional data. When the intervals between time points are long, there tends to be more missing data in longitudinal studies, often systematically related to variables inside or outside the study. For instance, in longitudinal

studies of school achievement, at time 2 and beyond, data are more likely to be missing for low income students and low scoring students at time 1 because such students tend to change schools more frequently. Therefore, researchers might want to apply some model-based, multiple imputation technique before computation of the proximity measure. Model-based imputation may be necessary to account for the systematic nature of the missingness. The method of handling missing data needs to be described and justified.

21. Interpretation of Dimensions in Longitudinal Studies

Interpretability depends primarily on whether what is known about the change process is consistent with the change patterns represented by scale values. For instance, if change is thought to be monotonically increasing with time, then a dimension along which scale values do not increase with time would be implausible and uninterpretable.

In longitudinal applications, the zero point along a dimension can be set in different ways without loss of fit to the data, and the different ways of setting the zero point lead to different interpretations of the intercept term. For instance, the zero point can be set so that the scale value at time 1 equals 0 for every dimension, in which case, c_p becomes the model based estimate of initial (time 1) status for person p . This alternative seems most useful in studies of growth or studies of decay, that is, studies in which change increases/decreases monotonically over time. If the measured variable oscillates nonmonotonically over time, the zero point can be set so that the mean scale value equals zero along each dimension. In such cases, each dimension is interpreted as a pattern of oscillating change about person p 's "typical" level of performance represented by the intercept, c_p . In longitudinal applications, researchers need to explain and justify how the zero point along each dimension was set and the resulting interpretation of the intercept parameter.

While we have not discussed the correspondence weight parameters c_p in the model, they may sometimes enhance the interpretability and plausibility of a dimension. That is, if they enter into relations with external variables (e.g., individual differences in weights c_p are correlated with individual differences in ability, personality, or interests), then the correspondence weights may help understand the relation between individual differences in growth patterns and individual differences the external variable(s). This explanatory power of the weights enhances the interpretability of the dimensions. The interpretability of a dimension depends on the explanatory power of both the dimensions' scale values and its correspondence weights.

Longitudinal MDS analysis can be used as an exploratory analysis by itself or as a way of generating hypotheses about longitudinal patterns needed to account for change in a particular variable, hypotheses that will be subsequently tested in a second sample using structural equation modeling or other growth modeling techniques (see Chapter 13, this volume). Theory may be too imprecise to generate the detailed growth curve hypotheses required by confirmatory methods, and therefore a combination of theory and exploratory analyses may be necessary for hypothesis generation. In comparison to other methods, advantages of MDS include: (1) no need for *a priori* growth trend specifications (e.g., linear or quadratic), (2) simultaneous estimation of multiple growth curves, (3) estimation of a growth rate for each time interval in each change pattern (dimension), and (4) ready accommodation of unequally spaced time points. Whether used to study perceptions, cross-sectional variables, or repeated measures, interpretation of the dimensions and/or the configuration need not rely solely on subjective judgment. Associations of dimension scale values with external variables and cluster analytically defined groupings of stimuli (or variables) in the solution space are just two examples of procedures that can be used to more objectively confirm or disconfirm interpretations of the solution. Purely subjective interpretations of solutions are to be avoided.

References

- Borg, I., & Groenen, P. (2005). *Modern multidimensional scaling: Theory and applications*. New York: Springer.
- Commandeur, J. J. F., & Heiser, W. J. (1993). Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data. Tech. Rep. No. RR-93-03. Leiden, the Netherlands: Department of Data Theory, Leiden University.
- Cox, T. F., & Cox, M. A. A. (2001). *Multidimensional scaling* (2nd ed.). New York: Chapman Hall.
- Davison, M. L. (1983). *Multidimensional scaling*. New York: Wiley (reprinted by Krueger, 1991).
- De Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software*, 31, 1–30.
- Ding, C. (2005). Determining the significance of scale values from multidimensional scaling profile analysis using a resampling method. *Behavior Research Methods, Instruments, & Computers*, 37(1), 37–47.
- Ding, C., Davison, M. L., & Petersen, A. C. (2005). Multidimensional scaling analysis of growth and change. *Journal of Educational Measurement*, 42, 171–191.
- Kim, S.-K., Davison, M. L., & Frisby, C. L. (2007). Confirmatory factor analysis and profile analysis via multidimensional scaling (PAMS). *Multivariate Behavioral Research*, 42, 1–32.
- Kim, S.-K., Frisby, C. L., & Davison, M. L. (2004). Estimating cognitive profiles using profile analysis via multidimensional scaling (PAMS). *Multivariate Behavioral Research*, 39, 595–624.
- Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1–27.
- Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, 115–129.
- Lingoes, J. C. (1989). *Guttman-Lingoes nonmetric PC series manual*. Ann Arbor, MI: Mathesis Press.
- MacKay, D. B., & Zinnes, J. (2005). PROSCAL professional: A program for probabilistic scaling. Retrieved from www.proscal.com.
- Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika*, 42, 241–266.
- SAS Institute (1990). *SAS/Stat software: Change and enhancement (release 6.07 Program manual)*. Cary, NC: SAS Institute.
- SPSS (1999). *SPSS professional manual 10.0*. Chicago, IL: SPSS.
- Takane, Y., Young, F. W., & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42, 7–67.
- Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson psychoeducational battery—revised*. Allen, TX: DLM.

22

Multilevel Modeling

D. Betsy McCoach

In the social sciences, a large proportion of our data are hierarchical in nature. Examples of naturally occurring hierarchies include students nested within schools, patients nested within hospitals, workers nested within companies, husbands and wives nested within couple dyads, and observations within people. Most traditional statistical analyses assume that observations are independent of each other. The assumption of independence means that subjects' responses are not correlated with each other. This assumption might be reasonable when data are randomly sampled from a large population. However, when people are clustered within naturally occurring organizational units (e.g., schools, classrooms, hospitals, companies), the responses of people from the same cluster are likely to exhibit some degree of relatedness with each other, given that they were sampled from the same organizational unit. Multilevel modeling techniques allow researchers to adjust for and model this non-independence.

Multilevel models are also often referred to as *hierarchical linear models*, *mixed models*, *mixed effects models*, or *random effects models*. These terms are generally used interchangeably, although there are slight differences in the meanings of the terms. For instance, hierarchical linear model is a more circumscribed term than the others, as it assumes a normally distributed response variable. In contrast, mixed effects or random effects models are the most general terms, as they denote non-independence within a data set, but that non-independence does not necessarily need to be hierarchically nested. For instance, cross-classified random effects account for non-independence that is crossed, rather than nested. In longitudinal educational studies, students often change teachers or schools. Therefore, each student is crossed within a particular combination of teachers or schools. In such scenarios, students are *cross-classified* by two teachers, and in some cases, by two schools. This chapter focuses specifically on multilevel models, or models that exhibit a purely hierarchical data structure.

With clustered data, traditional statistical analyses that assume independence produce incorrect standard errors. In such a scenario, the estimates of the standard errors are smaller than they should be. Therefore, the Type I error rate is inflated for all inferential statistical tests that make the assumption of independence. In multilevel analyses, we explicitly estimate and model the degree of relatedness of observations within the same cluster, thereby correctly estimating the standard errors and eliminating the problem of inflated Type I error rates.

The advantages of multilevel modeling, however, are not merely statistical in nature. Multilevel analyses allow us to exploit the information contained in cluster samples to explain both the between- and within-cluster variability of an outcome variable of interest. These models allow us to use predictors at both the individual (or lowest) level (level 1), and the organizational (or higher) level (level 2) to explain the variance in the dependent variable. We can also allow the relation between an independent variable and the dependent variable to randomly vary across clusters. If we find that the impact of the independent variable on the dependent variable varies across clusters, we can try to explain the variability in this relation using cluster-level variables. For example, we can allow the relation between students' SES and achievement to vary by school. If we find that this relation does vary by school, we can try to explain that variability using school-level predictors, such as type of school, school SES, or average per-pupil expenditures. If a level 2 variable, such as average per-pupil expenditure, moderates the relation between a level 1 variable (SES) and the dependent variable (achievement), this is called a *cross-level interaction*. Thus, multilevel modeling allows us to simultaneously model the impact of both individual (or lower-level) and institutional (or higher-level) variables on the dependent variable of interest, as well as to model the cross-level interactions between higher-level and lower-level variables on the outcome of interest. Such analyses allow us to ask and answer far more nuanced questions than are possible within traditional regression analyses.

Finally, growth curve and other longitudinal analyses can be reframed as multilevel models, in which observations across time are nested within individuals. Using this framework, we can partition residual or error variances into those that are within-person and those that are between people. In such a scenario, between-person residual variance represents between-person variability in any randomly varying level 1 parameter of interest, such as the intercept (which is commonly centered to represent initial status in growth models) and the growth slope.

Contemporary expositions of multilevel modeling include textbooks by Raudenbush and Bryk (2002), Hox (2010), and Snijders and Bosker (2012), and an edited volume by O'Connell and McCoach (2008). Table 22.1 presents specific desiderata for applied studies that utilize multilevel modeling, and the remainder of this chapter is devoted to the explication of these desiderata.

Table 22.1 Desiderata for Hierarchical Linear Modeling.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. Model theory and variables included in the model are consistent with the purposes of the study and the research questions or study hypotheses.	I
2. The decision to include/exclude random effects should be justified theoretically. The number of random effects to include should be as realistic and yet as parsimonious as possible. If random effects are eliminated during the model-building process, this decision should be justified both empirically and theoretically.	M, R
3. Statistical model is presented, preferably using equations. Otherwise, minimally, the statistical model is described in enough verbal detail to be replicable by other researchers, and for the reader to determine the fixed effects and the random effects at each level for each model.	M
4. Sample size is specified at each level, and is sufficient for conducting the proposed analysis. Sampling strategy and mode(s) of data collection are identified and justified. If appropriate, weighting methods are described and justified.	M
5. Measurement of the outcome/response variable is described and justified. Measurement of all explanatory variables is described and justified; evidence of reliability and validity is provided.	M

(continued)

Table 22.1 (*continued*)

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
6. Scaling and centering of predictor variables are described and justified. Coding of all categorical predictors is fully described. Special attention must be paid to the centering/coding of all lower-level independent variables and to the implications of these centering decisions for interpretation of the model results.	M
7. Extent of missing data is clearly reported for all variables at all levels, and methods for accommodating missing data are described. The final analytical sample is described.	M, R
8. For longitudinal models, the shape of the growth trajectory is described, and the modeling of this trajectory is described and justified.	M, R
9. The software or program, including version number, used to run the models should be identified. Parameter estimation strategy (e.g., REML, ML) is identified and justified.	M, R
10. Assumptions of the model are described and checked. This may include discussions of normality, outliers, multicollinearity, homogeneity or heterogeneity of variances, and residual diagnostics.	M, R
11. The assumed error covariance structure should be described, and any plausible alternative error covariance structures should be described and tested. This is especially important for longitudinal models.	M, R
12. Descriptive statistics for variables at each level of the analysis should be reported. These should include means, standard deviations, and correlations.	R
13. The intraclass correlation coefficient for the unconditional model should be reported and interpreted.	R
14. Generally, multilevel models are built sequentially, using a series of models: an unconditional model, a random coefficients model (containing lower-level predictors), and a full model (containing predictors at all level of the analysis). This series of models is described.	M, R
15. The write-up includes a table that presents the results of the analysis. These results should include both fixed effect parameter estimates and variance component estimates.	R
16. Model fit issues are addressed. Deviance is reported for any estimated models. Additionally, other measures of model fit (e.g., AIC, BIC) are reported for all estimated models. Competing nested models are compared using the likelihood ratio/chi-square difference test.	R
17. Some description/summary of the final model's predictive ability should be provided. This could include a proportion reduction in variance at each level / proportion of variance accounted for at each level. In models with random slopes, proportion reduction in variance measures can be misleading, as the predictive ability of the model varies by cluster.	R
18. Some measure of effect size or practical importance should be reported for the targeted coefficients of interest.	R, D
19. Language used in the presentation and discussion of results appropriately reflects the study design. Causal language is not used except when justified through study design.	R, D

* I = Introduction, M = Methods, R = Results, D = Discussion.

1. Model/Theory Alignment

When using modeling techniques such as multilevel modeling, a coherent conceptual base should inform and guide the statistical analyses. The model theory and the variables included in the model need to be consistent with the purposes of the study and the research questions or study hypotheses. Of course, any study should be guided by a coherent theory. However, given that excluding an important potential confounder creates the potential for bias in the estimates, the temptation with

regression-type models is to try to add any variable that might be related to the outcome variable. While it is true that failing to include important potential confounders can create bias in the estimates of the effects of other variables, given the complexity of the error structure and the number of potential cross-level interactions, models that include large numbers of fixed and random effects can become unwieldy, difficult to interpret, and perhaps even impossible to estimate. Therefore, researchers should spend a great deal of time determining the variables for inclusion based on theory and relevant literature prior to undertaking the data analysis.

2. Random Effects

As in multiple regression (see Chapter 23, this volume), we estimate a within-cluster residual (r) that represents the deviation of a person's score from his or her predicted value. In a multilevel model, the intercept and the slopes for each of the level 1 variables can randomly vary across the level 2 units. In general, we allow the intercept to randomly vary across level 2 units. Therefore, we estimate a residual for each cluster (u_0). This is the deviation of a cluster's value from the overall intercept. It is this ability to partition variance into within-cluster variance and between-cluster variance that is the essence of the multilevel model. For simplicity, imagine a model in which there are no predictors. Each person's score on the dependent variable is composed of three elements: the overall mean (γ_{00}), the deviation of the cluster mean from the overall mean (u_{0j}), and the deviation of the person's score from his/her cluster mean (r_{ij}). The u_0 term allows us to model the dependence of observations from the same cluster because u_{0j} is the same for every student within school j (Raudenbush & Bryk, 2002). The u_0 term is referred to as a *random effect* for the intercept because we assume that the value of u_0 randomly varies across the level 2 units (clusters). We also assume that u_0 is normally distributed with a mean of 0 and a variance of τ_{00} .

Now, imagine a model in which there is one predictor at the lowest level. For this example, assume that we are predicting reading achievement (Y_{ij}) using socio-economic status (SES). We continue to allow the intercept to randomly vary across schools. However, now we can allow the SES slope to randomly vary across schools as well by including u_1 . By allowing the SES slope to randomly vary across schools, we are specifying a model in which the relation between SES and reading achievement is different for different schools. Therefore, in some schools, there could be no relation between students' SES and their reading achievement, whereas in other schools the relation between students' SES and their reading achievement could be quite strong. The set of equations for this model is

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j}(SES)_{ij} + r_{ij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \tag{1}$$

Generally speaking, in multilevel modeling, a fixed effect represents the average effect across an entire population and is expressed by the regression coefficient (Snijders, 2005). In contrast, a random effect varies randomly across the population of level 2 units and is estimated as a residual for each of level 2 units (Snijders, 2005). Multilevel techniques allow us to model, estimate, and test the variances (and covariances) for these random effects. The variances and covariances of the random effects are referred to as *variance components*. In the set of equations above (1), the γ terms are the *fixed effects* and the u terms are the *random effects*.

In a two-level model, the number of possible random effects is equal to the number of variables at level 1 plus 1 (the random effect for the intercept). Therefore, in a model that contains 10 different level 1 variables, there could be up to 11 random effects. The number of random effects included should be as

realistic and yet as parsimonious as possible. At first glance, it might seem desirable to try to allow the slopes for all level 1 variables to vary randomly across the level 2 clusters and then to eliminate empirically any random effects that are not statistically significant. However, Raudenbush and Bryk (2002, p. 256) cautioned against this practice: “If one overfits the model by specifying too many random level 1 coefficients, the variation is partitioned into many little pieces, none of which is of much significance.” Instead, researchers should make the decision to include/exclude random effects based on theoretical grounds, rather than blindly allowing all level 1 slopes to vary randomly across level 2 clusters.

Even when using theory as a guide, sometimes analysts make changes to the random portion of the model during the model-building process. If random effects are eliminated during the model building process, this decision should be justified both empirically and theoretically. One common reason for eliminating random effects is that the level 2 variables in the model are able to explain the between-cluster variability in the slopes. For example, imagine a model in which students’ SES is a positive predictor of reading achievement and the SES slope randomly varies across schools. However, when the level 2 model includes the percentage of students within the school who are eligible for free lunch, the between-school variability in the SES slope is greatly reduced, and it is no longer statistically significant. In such a scenario, the between-school variability in the relation between SES and reading achievement is explained by a school-level variable: the percentage of students within the school who are eligible for free lunch. If the fixed effect for this cross-level interaction is negative, schools with greater percentages of students who are eligible for free lunch have less positive SES/reading slopes. If the fixed effect for this cross-level interaction is positive, then schools with larger percentages of free lunch students have larger, more positive SES/reading achievement slopes. In such a scenario, the slope of vocabulary on reading achievement is neither fixed nor randomly varying. Instead, it systematically varies as a function of the two level 2 variables.

3. Presentation of the Statistical Model

It is important for readers to be able to understand and potentially replicate the reported multilevel analyses. Therefore, the full hierarchical linear model must be specified clearly within the Methods section. There are many decisions that a researcher must make when building a multilevel model. Are the slopes of the level 1 coefficients allowed to vary randomly across the level 2 units? Which cross-level interactions between level 1 variables and level 2 variables are specified? Given the complexity of most multilevel models, the clearest and easiest way to communicate the exact specification of the model is to present the statistical model using equations. The equations for the multilevel model can be presented in one of two ways: using separate equations for the level 1 and level 2 variables or using a combined model.

To illustrate the multilevel and combined specifications, imagine a model in which the researcher wants to predict the reading achievement scores for students nested within schools. The level 1 independent variable is socio-economic status (SES), and the effect of SES is assumed to vary randomly across schools. The level 1 intercept is also allowed to vary randomly across schools. The level 2 independent variable is percentage of students receiving free-lunch (FREELNCH), which serves as an indicator of School SES. The multilevel, multiple equation notation is:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}(FREELNCH)_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(FREELNCH)_j + u_{1j},\end{aligned}\tag{2}$$

The γ_{00} term is the intercept of the school intercepts, indicating the predicted value of reading achievement when all other variables in the model are held constant at 0. The γ_{01} term represents

the unit change in the predicted value of the intercept per unit change in the free lunch variable. The γ_{10} term is the intercept of the SES slope, indicating the relation between SES and achievement when FREELNCH = 0. Finally, γ_{11} is the cross-level interaction between FREELNCH and SES, indicating the degree to which the percentage of students within a school who are eligible for free lunch moderates the relation of SES with reading achievement. The u_{0j} term indicates that the intercept (β_{0j}) is allowed to vary randomly across schools. The u_{1j} term indicates that the slope of the SES variable (β_{1j}) is allowed to vary randomly across schools. The combined model is the same model and contains the same information as the multilevel, multiple equation notation. However, in the combined model, we substitute the expressions to the right of the equals sign for β_0 and β_1 . Thus, the combined notation for the same model would be:

$$Y_{ij} = \gamma_{00} + \gamma_{01}(FREELNCH) + u_{0j} + \gamma_{10}(SES)_{ij} + \gamma_{11}(FREELNCH)_j(SES)_{ij} + u_{1j}(SES)_{ij} + r_{ij}. \quad (3)$$

Generally, these terms are regrouped so that the fixed effects are in the beginning of the equation and the random effects are at the end of the equation; so, the standard combined form would be as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{01}(FREELNCH)_j + \gamma_{10}(SES)_{ij} + \gamma_{11}(FREELNCH)_j(SES)_{ij} + u_{0j} + u_{1j}(SES)_{ij} + r_{ij}. \quad (4)$$

In reality, the model that is estimated is the combined model. Users of SAS, Stata, R, and SPSS must specify the combined model, whereas users of the software package HLM (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004), for example, may use the multiple equation notation to estimate multilevel models. Thus, certain multilevel modelers prefer to use the combined notation while others prefer the multiple equation notation. Either convention is acceptable, as both sets of equations are equivalent and contain the same information.

There might be audiences who would be confused by the multilevel equations. In such a situation, a reasonable solution is to present the equations and then to explain them verbally within text. Occasionally, a researcher might present research findings to an audience who would be completely overwhelmed by the presentation of equations, and the editor may request that the equations be removed from the manuscript. In such a circumstance, the statistical model should be described in enough verbal detail to be replicable by other researchers based on the description. The reader should be able to determine the fixed effects and the random effects at each level and the cross-level interactions. Some models are so complex that describing them verbally might actually be more difficult than using equations. Even so, the author should make every effort to present his or her multilevel models both verbally and through the use of combined or multilevel equations.

4. Sample Size Issues

Issues related to sample size are critically important in hierarchical linear models. Further, sample size issues are complicated by the multilevel nature of the data. In a sense, there are two salient sample sizes in a two-level model. Consider an organizational model in which people (level 1) are nested within organizations (level 2). The number of individuals represents the level 1 sample size, and the number of organizations represents the level 2 sample size. The number of level 1 units divided by the number of level 2 units provides an estimate of the average cluster size (the average number of level 1 unit within each of the level 2 units.) In a longitudinal model, on the other hand, observations across time are nested within people. Therefore, the number of observations across time (and people) is the level 1 sample size, and the number of people is the level 2 sample size. For example, in a longitudinal model where 100 people are each measured across four time points, the level 1 sample size would be 400 (4×100), and the level 2 sample size would be 100.

Generally speaking, the overall sample size is less important than the number of level 2 units and average number of level 1 units within each of the level 2 units. Thus, it is important to report the sample size at each level. At a minimum, the researcher should report the number of level 1 units, the number of level 2 units, and the mean and standard deviation of the average cluster size. If there are a small number of clusters ($N < 50$), including a frequency table that shows the number of observations within each cluster can be useful (Ferron, Hogarty, Dedrick, Hess, Niles, & Kromrey, 2008). In addition, the researcher should identify and justify the sampling strategy and the mode of data collection. When using sampling weights, it is important to describe the method for weighting the data and justify the decision to use sample weights. For more information on the use of sampling weights, see Stapleton and Thomas (2008).

There are two important considerations related to sample size. First, the number of units at each level of analysis must be large enough to estimate the multilevel model. Second, the analysis should be adequately powered to detect the effect of interest.

The total sample size is far less important than the average sample size at each level. Of course, the number of level 1 units can vary greatly from cluster to cluster. The number of level 1 units within clusters places an upper limit on the number of slopes that can randomly vary across clusters. The maximum number of random effects that can be estimated is one fewer than the number of level 1 units within the level 2 clusters; however, it is better to far exceed that minimal criterion. Although small average numbers of level 1 units within level 2 clusters limit the number of random effects that a researcher can estimate, the number of level 2 units is the more important sample size to consider when conducting or evaluating multilevel analyses. The sample size must be large enough to produce estimates, and these estimates must be reasonably free from bias. Maximum likelihood estimation (see Desideratum 9) is a large-sample technique that provides asymptotically unbiased estimates. However, in multilevel modeling, having a large overall sample size is not sufficient. The number of clusters (or the sample size at the highest level) must be large enough to support the estimation technique and to produce relatively unbiased estimates of the parameters and standard errors. What is the minimum number of clusters for a multilevel analysis? Certainly, it seems clear that multilevel analyses require a bare minimum of 10 clusters (Snijders & Bosker, 1999). However, such small sample sizes at level 2 might still produce biased estimates. The number of clusters impacts the estimates of the variance components and the standard errors, as well as the parameter estimates themselves and their standard errors.

Maas and Hox (2005) conducted a series of simulation studies to determine the smallest level 2 sample size that would produce unbiased parameter estimates and standard errors. With only 10 level 2 units, the regression parameters and the level 1 variance components exhibited little bias. However, the level 2 variance components were overestimated by approximately 25%, and the standard errors for all parameter estimates were underestimated, leading to inflated Type I error rates for all statistical hypothesis tests. With at least 30 clusters, the parameter estimates for the regression slopes and both the level 1 and level 2 variance components tended to exhibit very little bias in samples. However, there were issues with the estimation of the standard errors, especially for the variance components. Although the standard errors for the fixed effects and the level 1 variance components seemed to exhibit reasonable coverage with as few as 30 clusters, the standard errors for the level 2 variance components tended to be underestimated when there were fewer than 100 clusters (Maas & Hox, 2005). This means that studies with small to moderate numbers of clusters might have a higher Type I error rate for the level 2 variance components, which could lead to concluding mistakenly that the between-group variance is more pronounced than it actually is. Therefore, while it is possible to produce unbiased estimates of the fixed effects with as few as 10 higher-level units, at least 30 clusters are required to produce unbiased estimates

of the variance components and at least 100 clusters are necessary to have reasonable estimates of the standard errors of the level 2 variance components. In conclusion, while it may be possible to estimate a model with as few as 10 clusters, models with at least 30 clusters should provide reasonable estimates of variance components and standard errors for the fixed effects. However, standard errors for higher-level variance components are likely to be underestimated in studies with small to moderate numbers of clusters and a model comparison approach should be used instead (see Desideratum 16).

Maas and Hox (2005) considered only normal dependent variables. Non-normal (discrete) dependent variables appear somewhat analogous. Using Monte Carlo simulation techniques, Paccagnella (2011) considered the effects of sample size, the number of quadrature points, and the magnitude of the intra-class correlation coefficient accuracy of parameter estimates and standard errors of estimates in logistic multilevel models using Gaussian quadrature estimation. His results generally parallel those of Maas and Hox (2005): as in the continuous case: estimates of the fixed effects are unbiased, even with small sample sizes. Generally, the standard errors of the fixed effect estimates were well estimated with as few as 50 clusters. The estimation of the variance components and their standard errors was more problematic. The estimates of the variance components are underestimated, although the magnitude of this bias decreased substantially as the number of clusters increased. The standard errors of the variance components also exhibited downward bias, even with very large sample sizes. Increasing the number of quadrature points helped to mitigate the bias variance components.

Of course, an additional sample size consideration involves statistical power and precision: the number of level 1 and level 2 units must be large enough to detect the effect of interest. In the simplest scenario, power in multilevel modeling is a function of the number of clusters, the number of units per cluster, the intraclass correlation coefficient (see Desideratum 13), and the effect size. Although increasing sample size at either level increases power, in general, increasing the number of clusters boosts statistical power much more than increasing the average number of units per cluster does. This effect is even more pronounced as the intraclass correlation increases. Several free software programs are available to conduct a priori power analyses for multilevel models. These include the Optimal Design software program (Spybrook, Raudenbush, Liu, Congdon, & Martinez, 2011; <http://hlmsoft.net/od>) and the Power Up software and program (Dong & Maynard, 2013; [www.causalevaluation.org](http://causalevaluation.org)).

5. Measurement Issues

As with any analysis, it is important to describe the scale of measurement of the outcome variable. Hierarchical linear models are appropriate for analyzing continuous, normally distributed outcome variables whereas hierarchical *generalized* linear models allow for the estimation of non-normal response variables (O'Connell, Goldstein, Rogers, & Peng, 2008; Raudenbush & Bryk, 2002).

In addition, the Methods section should include a description of the scale of measurement for all of the explanatory variables in the model. As with any statistical analysis, the researcher should provide evidence of reliability and validity of each of the variables in the model. Because multilevel modeling is a regression-based technique, the assumptions of linear regression models (aside from the assumption of independence, which applies at each level) continue to apply (see Chapter 23, this volume). One commonly overlooked and rarely satisfied assumption of linear regression is that the independent variables are measured with perfect reliability. When one or more predictor variables are measured with error, the regression coefficients are likely to be biased: such biases can result in misleading inferences. Therefore, it is especially important to provide evidence of reliability of scores for all of the continuous independent variables in the model.

6. Centering

In multilevel modeling, it is especially important to describe and justify the scaling and centering of all the predictor variables. Decisions about centering impact the interpretation of the parameter estimates. Centering decisions are especially important for the lower-level independent variables because the choice of centering at the lower level(s) impacts the interpretation of both the lower- and higher-level parameter estimates. For organizational models, the two main centering techniques for lower-level independent variables are *grand mean centering* and *group mean centering*. In grand mean centering, the overall mean of the variable is subtracted from all scores. Therefore, the new score captures a person's standing relative to the full sample. In group mean centering, the cluster mean is subtracted from the score for each person in that cluster. As such, the transformed score captures a person's standing relative to his or her cluster. Whereas grand mean centering is a simple transformation of the raw score, group mean centering is not. There is some debate within the multilevel literature about whether grand mean centering or group mean centering is preferable from a statistical point of view. However, most experts in multilevel modeling agree on three issues related to centering. First, the decision to use grand mean or group mean centering should be based on substantive reasons, not just statistical ones. For instance, if the primary research question involves understanding the impact of a level 2 variable on the dependent variable and the level 1 variables serve as control variables, grand mean centering may be the most appropriate choice. On the other hand, when level 1 variables are of primary research interest, group mean centering may be more appropriate. This is because group mean centering removes between cluster variation from the level 1 covariate and provides an estimate of the pooled within cluster variance (Enders & Tofghi, 2007). Second, it is important to explain the centering decision and procedures and to interpret the parameter estimates accordingly. Third, when using group mean centering, it is important to introduce an aggregate of the group mean centered variable (or a higher-level variable that measures the same construct) into the analysis. Without an aggregate or contextual variable at level 2, all of the information about the between-cluster variability is lost. See Enders and Tofghi (2007) for an excellent discussion of centering in organizational multilevel models.

In growth models, the time or age variable also needs to be centered so that the intercept represents an interpretable value. For linear growth models, the most common technique is to center time at initial status or age at the beginning of the study. When time is centered at initial status, then the intercept represents an individual's starting value. However, the time variable can be centered at any point in the data collection period. For certain research problems, analysts may prefer to center time at the final time point or at the middle of the data collection cycle. As Biesanz and colleagues stated:

The choice of where to place the origin of time has to be substantively driven. Because this choice determines that point in time at which individual differences will be examined for the lower order coefficients, the answer to which coding(s) of time to examine in detail lies with the researcher's specific substantive questions of interest.

(Biesanz, Deeb-Sossa, Papadakis, Bollen, & Curran, 2004, p. 37)

Again, various options for centering time are appropriate and interpretable; however, it is incumbent upon the researcher to describe the centering procedure, the rationale for selecting the procedure, and the correct interpretation of the parameters, given the chosen centering procedure.

In addition to describing the centering and scaling of continuous variables, it is also important to describe the coding of all categorical predictors. Researchers should use the same conventions that they would use when conducting multiple regression to code the categorical variables in their models. Thus, researchers should use dummy coding, weighted or unweighted effects coding, or

contrast coding for all categorical variables (see Cohen, Cohen, West, & Aiken, 2003, for an excellent discussion of coding for multiple regression analyses). The decision about the type of coding scheme should be conceptually driven and result in easily interpretable parameter estimates. Again, researchers should describe the chosen coding scheme used and explain the appropriate interpretations of the parameter estimates that result from such a coding scheme. Finally, researchers need to consider the necessity of and model all same-level interactions among categorical and/or continuous variables in the same manner as they would if they were conducting a multiple regression analysis. The interpretation of the same level interaction parameter estimates depends on the coding schemes used for the lower-order variables. Often it is easiest to describe such interactions visually, using figures. Alternatively, creating tables of prototypical predicted values for different types of participants may help to illustrate such interaction effects. See Aiken and West (1991) for an excellent discussion of creating and interpreting same-level interactions within a multiple regression framework.

7. Missing Data

The percentage of missing data should be reported for all variables at all levels, and the author should describe the methods used to address the issue of missing data. Missing data are a problem for any analysis. However, in multilevel modeling, dealing with missing data can be especially complex. First, in most commercial multilevel software programs, units with missing data on any of the covariates are eliminated from the analysis by default. This becomes especially problematic when higher level units have missing data on any covariates, as the deletion of one higher level unit could result in the loss of tens, hundreds, or even thousands of lower level units, depending on the within cluster sample size of that higher level unit. For example, any school with missing data on any of the school-level covariates (e.g., percentage of free lunch eligible students, average per pupil expenditures) is eliminated from the multilevel model. Thus, it is easy to see how even small amounts of missing data at the higher levels of analysis could drastically reduce the size of the sample as well as the generalizability of the results.

Several modern data techniques exist for dealing with the problem of missing data. *Multiple imputation* (MI; Rubin, 1987, 1996) and full maximum likelihood estimation (FIML) (Enders, 2010) are generally considered two of the best methods of dealing with missing data. The use of multiple imputation has become increasingly common. When using MI with clustered data, there is one important caveat: multiple imputation of either the dependent variable or lower-level covariates should take the clustered nature of the data into account (Black, Harel, & McCoach, 2011). Standard multiple imputation procedures assume that the observations are independent. Using normal theory/standard approach does not provide valid inferences about variance components with any amount of missing data, and it does not provide reasonable estimates for fixed effects with high rates of missing data (Black et al., 2011). Although listwise deletion is generally considered a less desirable method of dealing with missing data than multiple imputation, there is some evidence to suggest that listwise deletion outperforms standard multiple imputation in terms of recovering parameter estimates when the data are multilevel in nature in some conditions (Black et al., 2011), however more advanced approaches would be prefer (Hox, van Buuren, & Jolani, 2016). When describing the sample, the author should explicitly describe the amount of missing data and justify his or her method of handling missing data.

8. Fitting Growth Trajectories

Fitting longitudinal growth models using hierarchical linear modeling techniques is becoming increasingly popular. In such a model, observations across time (at level 1) are nested within people (at level 2). Then standard unconditional linear growth model is

$$\begin{aligned}
 Y_{it} &= \pi_{0i} + \pi_{1i}(TIME)_{it} + e_{it} \\
 \pi_{0i} &= \beta_{00} + r_{0i} \\
 \pi_{1i} &= \beta_{10} + r_{1i}
 \end{aligned} \tag{5}$$

The dependent variable (y_{it}) is the score for student i at time t , which is a function of the randomly varying intercept, π_{0i} (which is the predicted value of y_{it} when time=0), the randomly varying growth slope, π_{1i} (time_{it}), and time-specific individual error.

Both the slope and the intercept contain a subscript i , indicating that a separate slope and intercept are estimated for each person in the sample. The deviation of a particular observation from the model-predicted trajectory is captured in the error term, (e_{it}) which represents the within-person error associated with that individual's data at that time point. The pooled error variability within individuals' trajectories is estimated by the variance of e_{it} [$\text{var}(e_{it}) = \sigma^2$] (Raudenbush & Bryk, 2002), and this error variance is generally assumed to be constant across time.

Because the time slope, π_{1i} , enters the equation as a predictor of the outcome value at a given occasion, each participant can have his/her own unique data collection schedule. Therefore, multilevel models seamlessly handle time unstructured data. Centering the time variable around some meaningful value within the data collection period helps to ensure the interpretability of the intercept and the variance/covariance components. In longitudinal studies, time is often centered at the beginning of the study period so that the intercept represents the expected value at the beginning of the study. If age is used as the time variable, it is quite common to center at a particular age (for example, at age 6). This strategy has the added advantage of controlling for age in addition to centering time. For example, if we center time at age 6, then the intercept represents the model predicted score at age 6.

The level 2 equations model the average growth trajectory across people and can capture between-person differences in the model-implied growth trajectories based on level 2 (time invariant) covariates. The second level of the multilevel model specifies that the randomly varying intercept (π_{0i}) for each individual (i) is predicted by an overall intercept (β_{10}), the effects of any level 2 predictors on the intercept, and , the level 2 residual, which represents the difference between person i 's model predicted intercept (based on the overall intercept, β_{10} , and level 2 predictors) and his or her actual intercept. Likewise, the randomly varying linear growth slope (π_{1i}) for each individual (i) is predicted by an overall intercept (β_{10}), the effects of level 2 variables on the linear growth slope, and r_{1i} , the level 2 residual, which represents the difference between person i 's model predicted linear growth slope and his or her actual growth slope. The inclusion of the r_{0i} and r_{1i} in the level 2 equations allows for between-person variability in the intercepts and slopes. If the intercept is centered around initial status, then the variance in r_{0i} (τ_{00}) represents the between-person variability in initial status, or where people start. Likewise, the variance in r_{1i} (τ_{11}) represents the between-person variability in peoples' growth rates. The standardized covariance of the two level 2 residuals, τ_{01} from the unconditional linear growth model provides the correlation between initial status (or, more generally, the intercept) and growth.

As the name implies, a linear growth model assumes a straight-line growth trajectory. However, many growth processes do not follow a linear trajectory. Assuming a linear growth trajectory is very limiting, and it may result in a serious misspecification of the model. Other shapes are accommodated easily using a variety of strategies. These include estimating piecewise models, polynomial models, or other non-linear models, as well as introducing time-varying covariates (McCoach & Kaniskan, 2010; McCoach, Madura, Rambo, O'Connell, & Welsh, 2013; McCoach & Yu, 2016; Singer & Willett, 2003). Therefore, the researcher should empirically examine the shape of the individual and average growth trajectories descriptively prior to fitting any statistical models.

This information, in combination with the theory, can help guide decisions about the shape of the growth trajectory. When using multilevel modeling to fit longitudinal models, it is imperative that the researcher describe the shape of the growth trajectory, describe the level 1 model, and justify how the modeling procedure used at level 1 was able to capture the shapes of the growth trajectories for the sample.

9. Software and Parameter Estimation

The methods section should include the program or software package and version used to conduct the analysis. Many general purpose statistical software packages such as R, SPSS, SAS, and Stata have multilevel capabilities. In addition, specialized multilevel software programs such as HLM, MLwin and latent variable modeling programs such as Mplus and LISREL are popular choices for estimating multilevel models. All of these programs handle straightforward two-level models with normal response variables with ease. Where the programs differ is in their ability to handle more complicated models such as cross-classified models, three-level models, multilevel mediational models, or models with non-normal outcome variables. For an overview and comparison of these different software programs, see McCoach *et al.* (2018) and Roberts and McLeod (2008), as well as the reviews provided at the Multilevel Centre (www.bristol.ac.uk/cmm/learning/mmssoftware).

The two most common estimation techniques for hierarchical linear models with normal response variables are *maximum likelihood* (ML) and *restricted maximum likelihood* (REML). The two methods should produce similar results in terms of the fixed effects (regression parameters); however, they do produce different estimates of the variance components (Snijders & Bosker, 1999). In ML estimation the estimates of the variance and covariance components are conditional upon the point estimates of the fixed effects, whereas in REML they are not (Raudenbush & Bryk, 2002). Thus whereas REML estimates of variance-covariance components adjust for the uncertainty about the fixed effects, ML estimates do not. When estimating the variance components, REML takes “into account the loss of degrees of freedom resulting from the estimation of the regression parameters, whereas the ML method does not” (Snijders & Bosker, 1999, p. 56). When the number of clusters is very large, REML and ML results should produce similar estimates of the variance components. However, when the number of level 2 units is relatively small, the ML estimates of the variance components (τ_{qq}) are underestimated by a factor of $(J - F)/J$, where J is the number of level 2 units and F is the number of fixed effects. Therefore, REML is the preferred estimation strategy for models with relatively few level 2 units.

While REML may be preferable to ML for estimating the variance components, ML is often preferable to REML for testing model fit. The deviances of any two nested models that differ in terms of their fixed and/or random effects can be compared when using ML. In contrast, REML only allows for comparison of nested models that differ in their random effects (Snijders & Bosker, 1999, p. 89). In addition, information criteria, such as the AIC and BIC, should be based on the ML estimates of the deviance (see Desideratum 16 for information about deviance and model fit.)

For binary or ordinal response variables, the most common estimation techniques use quasi-likelihood estimators such as penalized quasi-likelihood (PQL) or Maximum likelihood approaches using Gauss–Hermite quadrature, adaptive quadrature, or Laplace algorithms (Bauer & Sterba, 2011). Although penalized quasi-likelihood tends to be faster, especially in models with large numbers of random effects, it does not produce a deviance statistic that can be used to compare competing models. However, both PQL and ML approaches using adaptive quadrature appear to perform well for binary and ordinal models (Bauer & Sterba, 2011).

10. Assumptions and Residual Analyses

As with any statistical analysis, it is important to check the assumptions of the model and to describe any violations of the assumptions. Many regression diagnostics for single-level models are applicable within the multilevel framework as well. These may include discussions of normality, linearity, outliers, multicollinearity, homogeneity or heterogeneity of variances, and residual diagnostics. However, because the regression model is operating on multiple levels, tests of the assumptions become a bit more complex and time consuming.

O'Connell, Yeomans-Maldonado, and McCoach (2016) recommend a three-stage approach to conducting residual analyses within a multilevel framework. Stage 1 focuses on the level 1 model and residuals, and includes checking assumptions such as homogeneity of level 1 variance, linearity, and normality using statistics and graphical displays including histograms and box plots, graphs of predicted means versus residuals, and normal probability plots. Stage 2 focuses on exploring the level 2 residuals using the level 1 model/residuals at level 2. In addition, stage 2 includes examining for influential, outlying or unusual level 2 units. Stage 3 examines the residuals at level 1 and level 2 using the level 2 model (O'Connell et al., 2016).

Most residual analyses can and should be conducted at each level of the analysis. For example, in a two-level model where, say, students are nested within schools, it is possible to have an outlier at the student level or at the school level. Researchers should carefully check the assumptions of their models, and they should include a short description of the procedures that they used to check their assumptions. In addition, they should describe any violations of the assumptions and the procedures that they used to rectify those violations (e.g., Were any outliers deleted? Were any variables transformed?).

11. Error Covariance Structure

The researcher should briefly describe the assumed error covariance structure. Any plausible alternative error covariance structures should be described and tested. Generally, the assumed error covariance structure is quite reasonable for organizational models. The simplest error structure for a two-level model with a random intercept is depicted in equation (6). In this matrix, there are as many rows and columns as there are level 1 units. In this example, the first six level 1 units are shown. The first three level 1 units belong to cluster 1 and the second three units belong to cluster 2. The total residual variance for each person in the model is the sum of the within cluster residual (σ^2) and the between-cluster residual (τ_{00}). The covariance between any two people who are members of the same cluster is accounted for by τ_{00} , the between cluster residual. Finally, the residual covariance between 2 members of two different clusters is assumed to be 0.

$$\begin{bmatrix} \sigma^2 + \tau_{00} & \tau_{00} & \tau_{00} & 0 & 0 & 0 \\ \tau_{00} & \sigma^2 + \tau_{00} & \tau_{00} & 0 & 0 & 0 \\ \tau_{00} & \tau_{00} & \sigma^2 + \tau_{00} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 + \tau_{00} & \tau_{00} & \tau_{00} \\ 0 & 0 & 0 & \tau_{00} & \sigma^2 + \tau_{00} & \tau_{00} \\ 0 & 0 & 0 & \tau_{00} & \tau_{00} & \sigma^2 + \tau_{00} \\ \dots & & & & & \end{bmatrix} \quad (6)$$

Describing the error covariance structure is especially important for longitudinal models. Models that fail to adequately account for the covariances among repeated measurements may result in

misleading inferences (Fitzmaurice, Laird, & Ware, 2004). On the other hand, when modeling these longitudinal covariances, the analyst's goal should be to "select the most parsimonious covariance structure that reasonably fits the data" (Wolfinger, 1996, p. 208).

The standard multilevel linear growth model imposes a very particular structure on the composite within-person/across time covariances (the composite of the covariances across waves). The structure is dependent on the number of random effects in the model. The maximum number of random effects that can be estimated in a repeated measures model is the number of waves of data minus 1. The standard multilevel linear growth model estimates a random effect for the intercept, a random effect for the linear growth slope, and a covariance between the intercept and the slope. Using the standard multilevel model, the model-implied variance-covariance matrix for a model with four waves of data is

$$\begin{bmatrix} \tau_{00} + \sigma^2 & & & \\ \tau_{00} + \tau_{01} & \tau_{00} + 2\tau_{01} + \tau_{11} + \sigma^2 & & \\ \tau_{00} + 2\tau_{01} & \tau_{00} + 3\tau_{01} + 2\tau_{11} & \tau_{00} + 4\tau_{01} + 4\tau_{11} + \sigma^2 & \\ \tau_{00} + 3\tau_{01} & \tau_{00} + 4\tau_{01} + 3\tau_{11} & \tau_{00} + 5\tau_{01} + 6\tau_{11} & \tau_{00} + 6\tau_{01} + 9\tau_{11} + \sigma^2 \end{bmatrix}. \quad (7)$$

Thus, all 10 unique elements of the variance covariance matrix for the four repeated measurements are estimated using four parameters: τ_{00} , the between person variance in the intercept, τ_{11} , the between person variance in the linear growth slope, τ_{01} , the covariance between the slope and the intercept, and σ^2 , the within person residual variance. Other options for estimating the covariance structure of the repeated measurements include fitting models with heterogeneous σ^2 across the time points, first order autoregressive models, first order moving average models, and unrestricted covariance matrices, to name a few. A complete treatment of this topic is beyond the scope of this chapter. However, researchers who are interested in learning more about covariance structures for repeated measures multilevel models should consult Singer and Willett (2003), and Wolfinger (1996).

12. Descriptive Statistics

As in any research study, it is important to provide the reader with tables of descriptive statistics. Minimally, the author should provide a table of means and standard deviations and sample sizes for all of the continuous level 1 variables used in the analysis as well as a table of means and standard deviations, and sample sizes for all of the continuous level 2 variables in the model. Dichotomous variables should be reported as proportions or percentages. In addition, the document should include a table of correlations corresponding to each level in the analysis. So, for a two-level model, one correlation matrix should detail the correlations among the level 1 variables, whereas another correlation table should provide the correlations among the level 2 variables, computed at the cluster level.

13. Intraclass Correlation Coefficient

The *intraclass correlation coefficient* (ICC) is the proportion of variance in the outcome variable that is between clusters, that is, the proportion of variance that can be explained by the clustering or grouping structure (Hox, 2002). Alternatively, one may interpret the ICC as the "expected correlation between any two randomly chosen units that are in the same group" (Hox, 2002, p. 15). The formula for the ICC is

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2}. \quad (8)$$

Where τ_{00} represents between cluster variance and σ^2 represents within cluster variance. The ICC is important to report because it indicates the degree of non-independence in the data. The higher the ICC, the more homogeneity there is within clusters (or the more heterogeneity there is between clusters). An ICC of 0 indicates independence of observations, and any ICC above 0 indicates some degree of dependence in the data. The smaller and more homogeneous the cluster is, the higher the expected ICC is (McCoach & Adelson, 2010). For example, in school effects research, ICCs typically range from .10 to .20. In dyadic research, on the other hand, ICCs above .50 are not uncommon.

The computation of the *design effect*, which indicates the degree to which the parameter estimates' standard errors are underestimated when assuming independence, utilizes the ICC (ρ) and the average number of units per cluster (\bar{n}_j):

$$\text{design effect} = \sqrt{1 + \rho(\bar{n}_j - 1)} \quad (9)$$

Generally, design effects below 2.0 are considered fairly small. However, keep in mind that even with a design effect as low as 1.5, the standard errors in a model that assumes independence of observations are underestimated by a factor of 1.5. Therefore, the Type I error rate is already noticeably inflated, even with such a small design effect (McCoach & Adelson, 2010).

14. Model Building

Generally, multilevel models are built sequentially, using a series of models. First, researchers estimate an *unconditional* (or *null*) *model*, which contains no predictors. The purpose of this model is to obtain estimates of the level 1 and level 2 variance components for comparison to later, more parameterized models and to estimate the ICC. The second model estimated is a *random coefficients model*, which contains the level 1 predictors (Raudenbush & Bryk, 2002). Depending on the researcher's theoretical framework as well as the sample size at level 1, the slopes for some of the level 1 predictors may be estimated as randomly varying across level 2 units, or they can be estimated as fixed across all level 2 units. Raudenbush and Bryk (2002, p. 256) cautioned against the "natural temptation," which "is to estimate a 'saturated' level 1 model . . . where all potential predictors are included with random slopes." Any level 1 slopes that do not have statistically significant variability across level 2 units should be fixed prior to conducting the full contextual analysis. The next model to be estimated is the *full contextual model*, which contains both level 1 and level 2 predictors. Level 2 predictors can be used to predict the intercept or the mean value of the dependent variable (when all of the level 1 variables are held constant at 0). Level 2 predictors also can help explain the variability of level 1 slopes across clusters. In such a scenario, the level 2 variable is used to predict the level 1 slopes, or the relationship of the level 1 predictor and the dependent variable across level 2 units. For example, imagine that SES is a level 1 predictor of math achievement. Sector, a level 2 variable that indicates whether a school is public or private, can be added as a predictor of the relation between SES and math achievement. This cross-level interaction indicates whether sector moderates the relationship between SES relates and math achievement. Finally, if any fixed or random effects are eliminated from the full model, a final contextual model should be estimated and compared to the prior model.

It is important to describe the process of building these sequential models. Analysts differ somewhat in their approaches to building multilevel models. Thus, authors must be sure to describe the model building process in enough detail that another analyst could replicate the entire model and decision sequence.

15. Tables

The Results section should include a table that presents the results of the analyses. If space allows, presenting the results of the entire series of models can be quite informative; however, minimally, the table should include the complete results from the final, full contextual model. These results should include the fixed effect parameter estimates, the random effect parameter estimates (the variances of the random effects), the standard errors for all parameter estimates, and tests of statistical significance for both the fixed and random effects. Also, the table may include covariances among the random effects. Reporting the covariances is expected in longitudinal models; it is less customary to report all covariances among level 2 residuals in organizational models.

16. Deviance and Model Fit

It is important to address model fit issues as part of the model building and testing process. The deviance compares the log-likelihood of the specified model to the log-likelihood of a saturated model that fits the sample data perfectly (Singer & Willett, 2003, p. 117). Specifically, deviance = $-2LL$, where LL is the log-likelihood of the current model minus the log-likelihood of the saturated model. Therefore, deviance is a measure of the badness of fit of a given model; it describes how much worse the specified model is than the best possible model. Deviance statistics cannot be interpreted directly since deviance is a function of sample size as well as the fit of the model.

When one model is a subset or special case of the other, the two models are said to be “nested” (e.g., Kline, 1998). In nested models, “the more complex model includes all of the parameters of the simpler model plus one or more additional parameters” (Raudenbush, Bryk, Cheong, & Congdon, 2000, pp. 80–81). When two models are nested, their deviance can be compared directly using the chi-square difference test. The deviance of the simpler model (D_1), which has p_1 degrees of freedom, minus the deviance of the more complex model (D_2), which has p_2 degrees of freedom ($p_2 < p_1$), provides the change in deviance ($\Delta D = D_1 - D_2$). As the number of parameters in a model increases, the deviance value decreases. In sufficiently large samples, the difference between the deviances of two hierarchically nested models is distributed as an approximate chi-square distribution with degrees of freedom equal to the difference in the number of parameters being estimated between the two models (e.g., de Leeuw, 2004).

In evaluating model fit using the chi-square difference test, the more parsimonious model is preferred, as long as it does not result in statistically significantly worse fit. In other words, if the model with the larger number of parameters fails to reduce the deviance by a substantial amount, the more parsimonious model is retained. However, when the change in deviance (ΔD) exceeds the critical value of chi-square with $p_2 - p_1$ degrees of freedom, then the additional parameters have resulted in statistically significantly improved model fit. In this scenario, the more complex model (i.e., with p_1 degrees of freedom) is favored.

Under ML estimation, the number of reported parameters includes the fixed effects (the y terms) as well as the variance/covariance components. When using REML, the number of reported parameters includes only the variance and covariance components. To compare two nested models that differ in their fixed effects, it is necessary to use ML estimation, not REML estimation. REML only allows for comparison of models that differ in terms of their random effects but have the same fixed effects. Because most programs use REML as the default method of estimation, it is important to remember to select ML estimation to use the deviance estimates to compare two nested models with different fixed effects (McCoach & Black, 2008).

17. Predictive Ability of the Model

In single-level regression models, an important determinant of the utility of the model is the proportion of variance explained by the model, or R^2 . Unfortunately, there is no exact multilevel analog

to the proportion of variance explained. Variance components exist at each level of the multilevel model; therefore, variance can be accounted for at each level of the multilevel model. In addition, in random coefficients models, the relation between an independent variable at level 1 and the dependent variable can vary as a function of the level 2 unit or cluster. Consequently, there is no constant proportion of variance in the dependent variable that is explained by the independent variable. Instead, the variance in the dependent variable that is explained by the independent variable varies by cluster. Finally, because the variance components are estimated using ML estimation, the estimation of the variance can differ slightly from model to model. Therefore, it is impossible to compute an R^2 value for the entire model. However, both Raudenbush and Bryk (2002) and Snijders and Bosker (2012) have proposed multilevel analogs to R^2 . In both cases, the authors provided two separate formulas: one to explain variance at level 1 and another to explain variance at level 2.

Perhaps the most common statistic used to estimate the variance explained is the *proportional reduction in variance* statistic (Raudenbush & Bryk, 2002). The proportional reduction in variance can be estimated for any variance component in the model. This statistic compares the variance in the more parameterized model to the variance in a simpler baseline model. To compute the proportional reduction in variance, subtract the remaining variance within the more parameterized model from the variance within a baseline model. Then divide this difference by the variance within the baseline model. That statistic is computed

$$\frac{\hat{\sigma}_b^2 - \hat{\sigma}_f^2}{\hat{\sigma}_b^2} \quad (10)$$

where $\hat{\sigma}_b^2$ is the estimated level 1 variance for the baseline model and $\hat{\sigma}_f^2$ is the estimated level 1 variance for the fitted model (Raudenbush & Bryk, 2002). At level 2, population variance components estimates are represented by $\hat{\tau}_{qq}$ and are given for the intercepts (β_{0j}) and each slope estimate ($\beta_{1j}, \beta_{2j}, \dots, \beta_{qj}$) that is allowed to randomly vary across clusters. The proportional reduction in the variance of a given slope, β_{qj} , is

$$\frac{\hat{\tau}_{qq_b} - \hat{\tau}_{qq_f}}{\hat{\tau}_{qq_b}} \quad (11)$$

where $\hat{\tau}_{qq_b}$ is the estimated variance of slope q in the base model and $\hat{\tau}_{qq_f}$ is the estimated variance of slope q in the fitted model.

It should be noted, however, that the proportion reduction in variance statistic does not behave like the familiar R^2 . First, the proportional reduction in variance statistic proposed by Raudenbush and Bryk (2002) represents a comparison of one model to another model, and as such it cannot be interpreted as an explanation of the absolute amount of variance in the dependent variable. In addition, the proportion reduction in variance statistic can be negative. This actually happens with some regularity when comparing the level 2 intercept variance of a completely null model (a random effects ANOVA model which includes no predictors at level 1 or level 2) to the level 2 intercept variance of a model that includes a group mean centered predictor at level 1. Finally, it is inappropriate to use this technique to compute the proportion reduction in variance for two models that differ in terms of the number of random slopes being estimated.

The second method of deriving a multilevel R^2 type statistic (Snijders & Bosker, 1994, 1999) produces measures of *proportional reduction in prediction error* for level 1 (the prediction of Y_{ij}) and level 2 (the prediction of $\bar{Y}_{.j}$). These statistics are only available for models that include random intercepts but not for random coefficients models, which include randomly varying slopes. Like the proportional reduction in variance static presented above, the proportional reduction in prediction error for

level 1 (the prediction of Y_{ij}) compares the amount of residual variance in the more parameterized model to a simpler baseline model. However, this formula uses the total estimated variance, $\hat{\sigma}^2 + \hat{\tau}_{00}$, to compare the two models. The rationale is that $\hat{\sigma}^2 + \hat{\tau}_{00}$ provides a reasonable estimate of the total sample variance of the outcome variable Y (Snijders & Bosker, 1994). Because $\hat{\sigma}^2 + \hat{\tau}_{00}$ is being used as a proxy for the total variance in the dependent variable, this formula is only appropriate for models without randomly varying slopes. Given a random intercepts only model, the prediction error for individual outcomes (Y_{ij}) is equal to the sum of the level 1 and level 2 variance components, $\hat{\sigma}^2 + \hat{\tau}_{00}$.

The proportional reduction of prediction error at level 1 compares the total residual variance of a fitted (or more parameterized) model, f , to that of a baseline (or less parameterized) model, b . The formula for R_1^2 is

$$R_1^2 = 1 - \frac{(\hat{\sigma}^2 + \hat{\tau}_{00})_f}{(\hat{\sigma}^2 + \hat{\tau}_{00})_b} \quad (12)$$

Where the fraction's numerator is the prediction error for the fitted model and the fraction's denominator is the prediction error for the baseline model.

With respect to level 2, Snijders and Bosker's (1999, p. 103) explained proportion of variance at level 2 is the proportional reduction in the mean squared prediction error for the cluster mean " $\bar{Y}_{.j}$ for a randomly drawn level-two unit j ." The prediction error for the group mean is

$$\frac{\hat{\sigma}^2}{n_j} + \hat{\tau}_{00} \quad (13)$$

Thus, the level 2 proportional reduction in the prediction error, R_2^2 , is

$$R_2^2 = 1 - \frac{\left(\frac{\hat{\sigma}^2}{n_j} + \hat{\tau}_{00} \right)_f}{\left(\frac{\hat{\sigma}^2}{n_j} + \hat{\tau}_{00} \right)_b} \quad (14)$$

Where the fraction's numerator is the prediction error variance for the fitted model and the fraction's denominator is the prediction error variance for the baseline model. In this case \bar{n}_j , is a representative value for average group size.

The various multilevel R^2 -type statistics described above provide heuristics to compare models in terms of their ability to "explain variance." However, it is important to remember their shortcomings. First, these estimates do not provide unequivocal estimates of the variance explained by a model. Instead, they compare two models in terms of their ability to reduce some type of variance at one of the levels of the hierarchy. Second, when a model contains random slopes, R^2 does not have a unique definition (Hox, 1998; Kreft, deLeeuw, & Aiken, 1995). The relation between the level 1 predictor and the dependent variable varies across level 2 units, and the level 2 variance estimate is not constant in these models (Snijders & Bosker, 1999). Therefore, the notion of a unitary proportion of variance explained ceases to exist. Finally, these statistics can produce negative estimates, which provides a clear indication that they are not actually proportions of variance explained. However, even given these shortcomings, multilevel R^2 analogs do help researchers to compare predictive ability of various multilevel models. Therefore, they should be reported within the Results section of a multilevel paper. When reporting their R^2 results, researchers should be sure to specify whether

they used Raudenbush and Bryk's (2002) or Snijders and Bosker's (1999) method to compute these proportional reduction in variance estimates, and they also should clearly specify which model they used as the baseline model and which model they used as the fitted (or more parameterized model) for each of their computations.

18. Effect Size

As with any statistical analyses, it is important to report effect size measures for multilevel models. The R^2 analogs described above can help researchers and readers to determine the impact that a variable or a set of variables has on a model. In addition, researchers can compute Cohen's d -type effect sizes to describe the mean differences among groups (see Chapter 6, this volume). To calculate the equivalent of Cohen's d for a group-randomized study (where the treatment variable occurs at level 2), use the following formula:

$$\delta = \frac{\hat{\gamma}_{01}}{\sqrt{\hat{\sigma}^2 + \hat{\tau}_{00}}} \quad (15)$$

(Spybrook et al., 2011). Assuming the two groups have been coded as 0/1 or $-.5/.5$ (or any centering that retains a one point difference between the two groups), the numerator of the formula represents the difference between the treatment and control groups. The denominator utilizes the σ^2 and τ_{00} from the unconditional model. In the unconditional model, the total variance in the dependent variable is divided into two components: the between-cluster variance, τ_{00} , and the within-cluster variance, γ_{01} .

To facilitate understanding among readers, researchers should consider including figures that illustrate cross-level interactions among variables. Just as plotting same level interactions facilitates an understanding of interaction effects (Aiken & West, 1991), similar visual graphics of interactions between two variables at different levels of the data hierarchy can effectively display cross-level moderation. In addition, researchers should include predicted values for prototypical participants. These predicted values also can help the reader to make sense of the magnitude of the effects that are being reported. Thus, they serve as a form of "unstandardized" effect size.

19. Causal Claims

Multilevel modeling solves certain statistical issues that arise from non-independent or clustered data, and it allows for more nuanced analyses of variables that occur at different levels of the hierarchy. However, any causal claims that can be made from a multilevel analysis are determined by the strength of the research design. As Kelloway (1995, p. 216) stated, "No amount of sophisticated analyses can strengthen the inference obtainable from a weak design." It is common to refer to "effects" in multilevel modeling. In fact, the entire lexicon of the technique is replete with references to fixed effects, random effects, cross-level interaction effects, and so forth. However, none of these "effects" should ever be interpreted as indicative of causation or a causal mechanism except under certain randomized designs. When writing the Results and Discussion sections of a multilevel article, researchers should choose their language carefully so as not to imply causal claims that cannot be substantiated or defended given the design of the study.

References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
 Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods*, 16, 373–390.

- Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A., & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods*, 9, 30–52.
- Black, A. C., Harel, O., & McCoach, D. B. (2011). Missing data techniques for multilevel data: Implications of model misspecification. *Journal of Applied Statistics*, 38(9), 1845–1865.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- De Leeuw, J. (2004). Multilevel analysis: Techniques and applications (book review). *Journal of Educational Measurement*, 41, 73–77.
- Dong, N., & Maynard, R. A. (2013). *PowerUp!*: A tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs. *Journal of Research on Educational Effectiveness*, 6(1), 24–67.
- Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.
- Enders, C. K., & Tofghi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12, 121–138.
- Ferron, J. M., Hogarty, K. Y., Dedrick, R. F., Hess, M. R., Niles, J. D., & Kromrey, J. D. (2008). Reporting results from multilevel analyses. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 391–426). Charlotte, NC: Information Age Publishing.
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley-Interscience.
- Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). New York: Springer Verlag.
- Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Hox, J. J., van Buuren, S., & Jolani, S. (2016). Incomplete multilevel data. In J. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications* (pp. 39–62). Charlotte, NC: Information Age.
- Kelloway, E. K. (1995). Structural equation modeling in perspective. *Journal of Organizational Behavior*, 16, 215–224.
- Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: The Guilford Press.
- Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, 30, 1–21.
- Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1(3), 86–92.
- McCoach, D. B., & Adelson, J. (2010). Dealing with dependence (Part I): Understanding the effects of Clustered Data. *Gifted Child Quarterly*, 54, 152–155.
- McCoach, D. B., & Black, A. C. (2008). Assessing model adequacy. In Ann A. O'Connell & D. Betsy McCoach (Eds.), *Multilevel modeling of educational data* (pp. 245–272). Charlotte, NC: Information Age Publishing.
- McCoach, D. B., & Kaniskan, B. (2010). Using time-varying covariates in multilevel growth models. *Frontiers in Quantitative Psychology and Measurement*, 1, 17.
- McCoach, D. B., Madura, J., Rambo, K., O'Connell, A. A., & Welsh, M. (2013). Longitudinal data analysis. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 199–230). Rotterdam: Sense Publishers.
- McCoach, D. B., Rifenbark, G., Newton, S. D., Li, X., Kooken, J., Yomtov, D., Gambino, A., & Bellara, A. (2018). Does the package matter? A Comparison of Five Common Multilevel Modeling Software Packages. *Journal of Educational and Behavioral Statistics*, 43, 594–627.
- McCoach, D. B., & Yu, H. H. (2016). Using Individual Growth Curves to Model Reading Fluency. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications* (pp. 269–308). New York: Springer.
- O'Connell, A. A., Goldstein, J., Rogers, H. J., & Peng, C. Y. J. (2008). Multilevel logistic models for dichotomous and ordinal data. In A. A. O'Connell & D. B. McCoach (Eds.) *Multilevel modeling of educational data*. (pp. 199–244). Charlotte, NC: Information Age Publishing.
- O'Connell, A. A., & McCoach, D. B. (Eds.) (2008). *Multilevel modeling of educational data*. Charlotte, NC: Information Age Publishing.
- O'Connell, A. A., Yeomans-Maldonado, G., & McCoach, D. B. (2016). Residual diagnostics and model assessment in a multilevel framework: Recommendations toward best practice. In J. R. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research* (pp. 97–135). Charlotte, NC: Information Age.
- Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models: New simulation results. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 7(3), 111–120.
- Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models* (2nd ed.). Newbury Park, CA: Sage.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T., Jr. (2000). *HLM 5: Hierarchical linear and nonlinear modeling*. Statistical software manual. Skokie, IL: Scientific Software International.
- Raudenbush, S., Bryk, A., Cheong, Y. F., Congdon, R., & du Toit, M. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.
- Roberts, J. K., & McLeod, P. (2008). Software options for multilevel models. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 427–467). Charlotte, NC: Information Age Publishing.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley & Sons.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473–489.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford.
- Snijders, T. A. B. (2005). Fixed and random effects. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (vol. 2, pp. 664–665). New York: Wiley.

- Snijders, T., & Bosker, R. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, 22, 342–363.
- Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.
- Snijders, T., & Bosker, R. (2012). *Multilevel analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Spybrook, J., Raudenbush, S. W., Liu, X., Congdon, R., & Martinez, A. (2011). Optimal design for longitudinal and multilevel research. V1.77 [computer software]. Retrieved February 20, 2016 from <http://hlmssoft.net/od/>
- Stapleton, L. M., & Thomas, S. L. (2008). Sources and issues in the use of national datasets for pedagogy and research. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel analysis of educational data* (pp. 11–57). Charlotte, NC: Information Age Publishing.
- Wolfinger, R. D. (1996). Heterogeneous variance covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, 1, 205–230.

23

Multiple Regression

Ken Kelley and Scott E. Maxwell

Multiple regression has been described as a general data analytic system (e.g., Cohen, 1968), primarily because many commonly used statistical models can be regarded as its special cases (e.g., single-sample t -test, two-independent samples t -test, one-way analysis of variance), the independent variables can be categorical (e.g., groups) or quantitative (e.g., level of treatment), and the model can be used for observational or experimental studies. Furthermore, many advanced models have multiple regression as a special case (e.g., path analysis, structural equation modeling, multilevel models, analysis of covariance). The ubiquity of multiple regression makes this model one of the most important and widely used statistical methods in social science research. In general, the idea of the multiple regression model is to relate a set of *regressor* (*independent* or *predictor*) variables to a *criterion* (*dependent* or *outcome*) variable, for purposes of explanation and/or prediction, with an equation linear in its parameters. More formally, the population multiple regression model is given as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \varepsilon_i, \quad (1)$$

where β_0 is the population intercept, β_k is the population regression coefficient for the k th regressor ($k = 1, \dots, K$), X_{ki} is the k th regressor for the i th individual ($i = 1, \dots, N$), and ε_i is the error for the i th individual, generally assumed to be normally distributed with mean 0 and population variance σ^2_ε . The intercept is the model-implied expected value of Y when each of the K X variables are at values of zero. The intercept may have a meaningful substantive interpretation, such as when the regressor variables are centered around 0 so that the intercept represents the grand mean on the outcome or when the regressor variables are dummy variables and the intercept thus represents the expected value of the outcome for the referent group, otherwise it serves as a scalar so that the sum of the squared errors can be minimized. For contemporary treatments of multiple regression applied to a wide variety of examples, we recommend Cohen, Cohen, West, and Aiken (2003), Pedhazur (1997), Harrell (2001), Fox (2008), Rencher and Schaalje (2008), Gelman and Hill (2007), and Muller and Fetterman (2002). Specific desiderata for applied studies that utilize multiple regression are presented in Table 23.1 and explicated subsequently.

Table 23.1 Desiderata for Multiple Regression.

<i>Desideratum</i>	<i>Manuscript Section</i>
1. The goals of the research and how multiple regression (MR) can be useful are explicitly addressed.	I
2. The inclusion of each of the independent variables, whether confirmatory or exploratory in nature, should be justified on theoretical and/or practical grounds.	I
3. Each criterion and regressor variable should be described in detail (e.g., scales of measurement, coding scheme, reliability) to convey how the MR model should be interpreted.	M
4. Specific procedures for the computation and interpretation of effect sizes are delineated.	M
5. Assumptions underlying the MR analyses and resulting inference are explicitly addressed.	M
6. Variable selection techniques are justified.	M
7. Sample sizes for all analyses are justified in terms of power, accuracy, and reproducibility of results.	M
8. Methods of dealing with missing data are addressed.	M
9. For models examining moderation, issues of interpretation, role of centering, and visualization are addressed.	R
10. For models examining mediation, issues of interpretation and limitations due to cross sectional designs are addressed.	R
11. Visual examination of data is addressed in order to assess model appropriateness and assumptions.	R
12. Measurement error in predictor and/or outcome variables is addressed.	D
13. Potential limitations of multiple regression in the current applied research context are explicitly stated.	D
14. Alternatives to the MR model are given.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

1. Research Goals

Standard textbook treatments of multiple regression often emphasize that multiple regression can be used for prediction or explanation. Depending on the goals of the researcher, prediction, explanation, or both might be desired. Although the multiple regression model itself is exactly the same in both cases (i.e., Equation (1) does not change based on the goal), the distinction is nevertheless important because different statistical considerations arise for the two purposes. To clearly communicate the purpose of the study, it is important for authors to be clear about whether their purpose in using multiple regression is prediction, explanation, or both.

The ultimate goal of explanation is to identify the causes of the outcome variable Y . Under ideal conditions, multiple regression can identify causal effects by assessing the value of regression coefficients: when the coefficients are non-zero in the population causality may be a possibility. To understand how a regression coefficient can potentially reflect a causal effect, we need to say what a regression coefficient represents. For example, when the model is correctly specified, the coefficient β_k for X_k reflects the linear relation between Y and X_k at a fixed value of all other regressors included in the model. In this sense the regression coefficient for X_k is a measure of the extent to which X_k and Y are linearly related when all other regressors in the model are held constant. Because the other regressors are held constant, any association between X_k and Y cannot be attributed to the other regressors. Thus, it is tempting to conclude that β_k reflects the extent to which X_k causes Y , in which case we have at least partly succeeded in explaining variation in Y . In fact, this reasoning is sometimes correct, but only under a set of restrictive conditions (temporal precedence, relationship,

and nonspuriousness; see, e.g., Kenny, 1979). Unfortunately, it can often be difficult to justify these conditions unequivocally except in randomized experiments.

Predicting the value of a criterion variable given one or more regressors is another reason why multiple regression is commonly used, especially in applied research. For example, a researcher might use multiple regression to predict how well pre-kindergarten children will be able to read at the end of first grade. The researcher would use historical data (often called *training data*) containing scores on reading at the end of first grade as well as scores on a number of possible regressors. Multiple regression could then be used to create a model in which the value of the criterion is predicted based on one or more of the regressors. A benefit of prediction is that the parameter estimates (i.e., the regression coefficients) obtained from the training data can be used to predict the value of an unknown (or yet to occur) criterion variable *Y* based on the complete set of regressors used in the training data. There are many cases in which it is desirable to predict a criterion variable when it is as yet unknown (e.g., college grade point average or reading ability at the end of first grade) from a set of known regressors (e.g., SAT scores or pre-kindergarten measures of cognitive functioning). The ultimate goal is often selection, as in the college example, but can also be identifying at-risk individuals who might benefit from a relevant intervention.

Although we believe that recognizing the difference between explanation and prediction is critically important when considering the parameters of interest in the model, there need not be such a rigid dichotomy between the two goals. In studies seeking to explain relations there can be prediction, and in studies that seek a way to predict there can be attempts at explanation. Pedhazur (1997, p. 196) described predictive research having as its main emphasis “practical applications,” whereas in explanatory research the main emphasis is “understanding phenomena.” Huberty (2003) provided a discussion of the similarities and differences in research goals and reporting strategies when interest is primarily in prediction or explanation.

Statistical inference is important when a desire exists to generalize information obtained in a sample to the population from which the sample was drawn. Inference can be of two forms, confidence interval formation for the population effect sizes of interest and/or hypothesis testing for effect sizes. For purely predictive purposes, inferential procedures are not strictly necessary, but nevertheless provide information about the population of interest.

2. Justification of Regressors

Multiple regression can be applied along a continuum of research approaches anchored by *confirmatory* and *exploratory* research. The confirmatory anchor seems to best correspond to a well-defined research question with a few theoretically justified variables, whereas the exploratory anchor corresponds to a diffuse research question with many variables included in one or more different analyses, not necessarily with explicit theoretical justification. Both confirmatory and exploratory analyses are beneficial, but care must be taken so that an exploratory analysis is not presented as if it were a confirmatory analysis. Provided the assumptions of the model are satisfied in the context of confirmatory studies, the probability values (i.e., the *p*-values) from null hypothesis significance testing and confidence interval coverages associated with the different effect sizes are meaningful. However, because exploratory analyses generally consist of systematic testing and retesting until settling on a model that is satisfactory enough, the process of testing and then retesting renders the probability values and confidence interval coverages associated with the effect sizes as approximate at best, and completely inappropriate at worst. For example, testing many models with the aim of obtaining *p*-values for coefficients of interest less than, say, .05, leads to *p*-values that are too heavily based on the characteristics of the sample rather than a test of a well-specified question. Testing null hypotheses of different models on the same data set will result in capitalization on chance and

more Type I errors will be realized than the Type I error rate specified by the researcher (e.g., .05). That is to say, when exploratory analyses are treated as if there were confirmatory, properties of the p -values will not be the same as if the analysis was truly confirmatory. Nevertheless, findings from such exploratory studies often provide a useful starting point for future confirmatory research but it needs to be clear to the reader how the analysis was conducted and what other analyses were attempted. Readers, such as reviewers, are rightly skeptical of drawing important conclusions from studies in which many models were evaluated and only the significant findings presented.

More formally, the reason probability values and confidence interval coverages are not correct in exploratory analyses in which multiple models are evaluated is because of what is known as the *multiplicity problem*. The multiplicity problem describes the problem of multiple statistical tests being performed, where the effect sizes with small p -values are selected for inclusion in the presented statistical model. An implication of the multiplicity problem is that the obtained p -values are suspect, due to the sheer number of null hypothesis significance tests conducted. When many null hypothesis significance tests are conducted, even when all the null hypotheses are true, there is a high probability of finding some small p -values by chance. Thus, because of the suspect p -values and the associated confidence interval coverages associated with statistical inference in exploratory studies, it should be made clear if the study was confirmatory in nature or exploratory. In particular, exploratory approaches sometimes effectively are based on an informal variation of a formal variable selection method (such as stepwise regression, to be discussed in Desideratum 6), which may be fine for prediction but raises serious concerns about the meaningfulness of any claims regarding explanation. That is, some researchers reject the idea of stepwise regression, but themselves perform a more intuitive version of stepwise regression where many models are fitted, even when their purpose is explanation.

3. Descriptions of Criterion and Regressor Variables

A statistical model in and of itself is not very useful unless the variables in the model are understood in their appropriate context and have been discussed in enough detail to convey an understanding of the information they contribute to the research question. At a minimum, means and the covariance matrix or the correlation matrix (with accompanying standard deviations) should be provided for all variables used in the analysis. Furthermore, the type of variable (e.g., categorical or continuous) and the range over which values of the scale can vary (i.e., the limits of the scale) should be discussed. When categorical variables (e.g., grouping variables) are used, the coding scheme should be explicitly discussed. Without an explanation of the coding scheme, the estimated model parameters cannot be readily interpreted by others (e.g., for the "Sex" variable females are coded as 0 and males 1, females as 1 and males 0, or females -1 and males 1, etc.). Continuous variables should almost never be dichotomized (or polytomized more generally) but should instead be left in their continuous form in order to preserve as much information in the variable as possible. Examples of situations where it may sometimes be reasonable to polytomize continuous variables is when there are clear types or taxa of individuals or when the distribution of a count variable is highly skewed (MacCallum, Zhang, Preacher, & Rucker, 2002). It is clear, however, that median splits, a commonly used procedure for dichotomizing continuous data, is essentially never statistically justified. Where appropriate, the reliability and validity evidence for each of the variables should be provided (see Desideratum 12); more information is available in Chapter 29 of this volume.

4. Effect Sizes

As has been discussed a great deal in the methodological literature, effect sizes and their corresponding confidence intervals are widely recommended and should almost always be reported (e.g.,

Wilkinson & APA Task Force on Task Force on Statistical Inference, 1999; see also Chapter 6, this volume). In multiple regression, like many other statistical models, there are two types of effect sizes: *omnibus* and *targeted*.

The most widely used omnibus effect size in multiple regression, and one of the most common in social science research in general, is the squared multiple correlation coefficient, whose population value is denoted P^2 (rho squared). The value of P^2 quantifies the proportion of variance in Y that can be accounted for by the K regressor variables. The typical estimate of P^2 , R^2 , is positively biased. Although confidence intervals and significance tests for P^2 are based on R^2 , the adjusted value of R^2 , denoted R_A^2 , should also be reported and used as the best estimate of P^2 . The typical adjusted estimate (e.g., Cohen et al., 2003; Harrell, 2001) is given as

$$R_A^2 = \max \left\{ 0, \left[1 - (1 - R^2) \left(\frac{N-1}{N-K-1} \right) \right] \right\}, \quad (2)$$

where $\max\{\cdot, \cdot\}$ implies that the larger of the two values is taken. Most statistical programs will give both R^2 and R_A^2 .

Darlington (1968) explained that the adjustment shown in Equation (2) (developed by Ezekiel, 1930) will tend to overestimate the population validity of the sample regression equation. The idea here is that the adjustment estimates the population validity of the population regression equation. In other words, if the population regression coefficients were known, what proportion of the variance in Y would this equation explain in the population? This makes sense when the goal is explanation, because one purpose here is to estimate the extent to which the regressors explain the variance in Y . However, this makes less sense when the goal is prediction, because in this context the sample regression equation derived in the training sample will be used to make predictions in a new sample. The key point is that the regression coefficients to be used for prediction are the values obtained in the training sample. However, these values will not be exactly the same as the optimal population values, thus lowering the resultant R^2 to some extent. For this reason, in the context of prediction, the population parameter of most interest is sometimes referred to as the population cross-validity, P_C , or the squared population cross-validity, ρ_c^2 . Raju, Bilgic, Edwards, and Fleer (1999) described a variety of estimators of the population cross-validity and recommended an adjustment developed by Burkett (1964):

$$R_C = \frac{NR^2 - K}{R(N - K)}. \quad (3)$$

Although these omnibus effect size estimates are beneficial, an observed effect size is simply a point estimate that might differ considerably from the population value it estimates. Confidence intervals should be reported for any estimate that is itself deemed important enough to report. Confidence intervals for P^2 are not straightforward to construct and the appropriate confidence interval depends on whether or not regressors are regarded as fixed or random. Steiger (2004; see also Steiger & Fouladi, 1992), Algina and Olejnik (2000), and Kelley (2007), discussed methods of confidence interval construction and provided software solutions to implement such intervals.

Researchers should consider the squared semi-partial correlation coefficient, which is a targeted effect that describes the change in R^2 when the k th regressor is added to the multiple regression model that already contains the other $K - 1$ regressors. Thus, the squared semi-partial correlation coefficient quantifies the proportion of variance of Y that is accounted for *uniquely* by a particular regressor in a model with other regressors. Such an effect size is useful when conveying the contribution of a regressor in a model with $K - 1$ other regressors. Squared semi-partial correlation

coefficients can also be used to quantify the proportion of variance of Y that is accounted for by a particular set of regressors instead of just a single regressor.

Regression coefficients come in two forms: *unstandardized* and *standardized*, both of which represent targeted effects, which may or may not be causal in nature. Unstandardized regression coefficients can be transformed into standardized regression coefficients by multiplying the unstandardized regression coefficient by the quantity $\frac{s_{X_k}}{s_Y}$, which removes the scale of X_k and Y , where s_\cdot denotes the standard deviation of the subscripted quantity. The process can be reversed (i.e., set a standardized regression coefficient on the unstandardized scale) by multiplying a standardized regression coefficient by $\frac{s_Y}{s_{X_k}}$. In general, either unstandardized or both unstandardized and standardized regression coefficients should be given, along with their corresponding confidence intervals. The k th regression coefficient quantifies the degree of linear relation between Y and X_k , while holding constant the remaining $K - 1$ regressors. Standardized regression coefficients are often an effective way of describing the effect of a regressor on the criterion variable when the scales of the measurements are not inherently meaningful. When standardized solutions are used in place of or in addition to their unstandardized counterparts, the measure of association is in terms of standard deviation units of the particular sample. For example, a standardized regression coefficient of .25 for X_k in a standardized solution implies that a 1 standard deviation unit difference in X_k is associated with a .25 standard deviation difference in Y in the same direction, holding constant all other regressors.

Confidence intervals for unstandardized regression coefficients are easy to obtain and formulas are available in essentially all modern regression books and can also be obtained with popular statistical software. However, confidence intervals for standardized regression coefficients require the use of noncentral t distributions and are more difficult to obtain (e.g., see Kelley & Maxwell, 2008, or Kelley, 2007, for a review and software solutions). In general, standardized regression coefficients are provided when there is a desire to remove the scaling of the measurement instrument so that each variable (regressors and criterion) has a mean of 0 and a standard deviation of 1. Standardized regression coefficients allow for relations to be framed in standard deviation units (as previously noted) and regression coefficients to be more directly comparable within an equation. That being said, there is no guarantee that the regressor with the largest regression coefficient is the “most important” independent variable in the equation (even when all variables are standardized). The meaning of “most important” might be different depending on the particular situation and goals of the study (Azen & Budescu, 2003).

5. Addressing Assumptions

Standard approaches to regression rely on ordinary least squares (OLS) to estimate model parameters. The OLS regression coefficients in multiple regression minimize the sum of squared deviations between the model implied scores, denoted \hat{Y}_i for the i th individual, and the observed scores (i.e., regression coefficients are chosen that minimize $\sum_{i=1}^N (Y_i - \hat{Y}_i)^2$). Estimation of the regression coefficients themselves does not strictly require any parametric assumptions. However, inference for the regression coefficients in the usual ways (hypothesis testing and confidence interval formation) does depend on validity of underlying assumptions. In particular, p -values and confidence intervals (i.e., inference) for regression coefficients from the regression model as specified in Equation (1) depend on four statistical assumptions: (a) errors (i.e., $e_i = Y_i - \hat{Y}_i$) follow a normal distribution; (b) error variance is homogeneous across all values of the regressors (*homoscedasticity*); (c) the entities (e.g., persons) from which observations are taken are independent of one another; and

(d) the relation between Y and the K regressors is linear. It is important to note that no distributional assumptions are made about the regressors, meaning that, for example, skewness in a predictor is not by itself a problem. Also, the model does not assume that regressors are measured without error, but as we will discuss later, results obtained using regressors measured with error may differ substantially from results obtained when regressors are measured perfectly, so measurement error in the regressors often becomes an important consideration.

Although the linearity assumption (assumption d above) is fundamental, it is often overlooked in discussions and applications of multiple regression. We agree with Gelman and Hill (2007, p. 46) that, “The most important mathematical assumption of the regression model is that its deterministic component is a linear function of the separate predictors.” This assumption is especially important because if this assumption is not valid, then the regression coefficients in the model do not accurately reflect the relation between Y and X_k at a fixed value of the other regressors. As a result, the regression model might fail to hold the other regressors constant in attempting to estimate the relation between Y and a specific X variable. If linearity does not hold, then the model as specified in Equation (1) may not be appropriate for inferences, as Equation (1) is necessarily linear in form. When linearity does not hold, there are essentially three strategies: (a) transform one or more variables (one or more X_k and/or Y) so that linearity in an additive model is a good approximation (e.g., $\sqrt{X_k}$ or X_k^2), (b) include an additional theoretically justified variable (e.g., X_k^2 in addition to X_k) that correlates with the outcome variable, in an attempt to explain some of the unaccounted for variability, and/or (c) fit a nonlinear regression model (e.g., a negative exponential, Gompertz, logistic) instead of the traditional linear multiple regression model (Seber & Wild, 1989).

6. Variable Selection Techniques Are Formally Justified

In many situations, more regressor variables are initially included in a model than are ultimately desirable in the final model to be presented for interpretation. The way in which the researcher arrives at the final model should be made explicit. There are four common ways of selecting variables to be included in the analysis: (a) all analyses are theory driven, (b) model comparisons are performed, (c) stepwise methods are used, or (d) a variety of exploratory models and methods are fitted.

In many ways, the ideal variable selection method is entirely theory driven and the regressors included are based on *a priori* theoretical arguments and/or previous literature. This method is ideal because a one-to-one mapping exists between the targeted nature of the research question and the targeted statistical analyses.

A model comparison approach (e.g., Maxwell, Delaney, & Kelley, 2018), in which the inclusion of one or more variables is evaluated against a more basic model, is often the most straightforward way to evaluate competing nested models (on the same set of data). The idea of the model comparison approach is to statistically compare nested models, where the models are compared most commonly in terms of R_k^2 and R_{K+M}^2 , where R_k^2 is the model based on the K regressors and R_{K+M}^2 is based on a richer model with an additional M regressors.

A special type of model comparison is implemented through what is often termed *hierarchical regression* (not to be confused with hierarchical linear modeling, HLM; see Chapter 22, this volume). In hierarchical regression, not only are the variables selected by the researcher, so too is the order in which they enter the model. At each step of the procedure, the variables previously included remain in the analysis. When hierarchical regressions are performed, a series of fitted models should be provided as part of the reported results that shows the estimated model improvement when comparing the richer models to the simpler models. The improvement is generally gauged in terms of the change in R^2 when a single regressor variable is added, which again is the squared semipartial correlation coefficient. It is also common to add a block of regressors in a hierarchical fashion.

In such situations the change in R^2 is still of interest, but there the additional variability accounted for is due to the block of regressors. For example, a researcher might add a block of control variables before adding one or more primary variables of interest.

When a large number of possible regressors exist, possibly for more than one criterion variable, data driven selection methods are sometimes used. Whenever data driven selection methods are used, a clear indication should be made that the study is not attempting to explain phenomena in a confirmatory fashion, but rather that the study is exploratory in nature. The type of data driven selection procedure performed (e.g., forward selection, backward elimination, all possible subsets), and the selection criteria (e.g., a statistically significant change in R^2 , or a change in R^2 of some specified magnitude, say .05) should be given. Also the particular computer program/package and its version should be provided, because different programs/packages and versions implement data driven selection procedures in different ways.

There are many methodological issues that can arise when implementing a data driven selection procedure. As Rencher and Pun (1980) illustrated, values of R^2 can be highly inflated and thus the obtained probability values can differ substantially from those reported as output in statistical software. When a large number of possible regressors exist in the context of a data driven selection procedure, a model that accounts for a statistically significant proportion of variance in Y can often be obtained even if the null hypothesis is true that all of the regression coefficients, less the intercept, are zero. Because of the multiplicity issue, as previously noted, fitting more than a single model can inflate the Type I error rate due to capitalization on chance.

Vittinghoff, Glidden, Shiboski, and McCulloch (2005) provided an especially interesting perspective on model building by distinguishing three different purposes for selecting predictors: (1) evaluating a regressor of primary interest in the context of other possibly relevant regressors, (2) identifying the important regressors of an outcome, and (3) prediction. They emphasized that issues involved in predictor selection differ according to the purpose of the analysis. For example, suppose that two regressors X_1 and X_2 are highly correlated with one another. When the goal is prediction, it will generally be desirable to include only one of these two regressors in the model, and it may make little difference in the accuracy of prediction which of the two is included. Ironically, however, including both of the regressors will often worsen prediction because any gain in bias reduction is more than offset by an increase in the variance of predicted values. On the other hand, suppose the goal is to explain the relation between X_1 and Y . Should X_2 be controlled for and thus included in the model? We agree with Vittinghoff et al. (2005) that this question cannot be answered simply from knowing that X_1 and X_2 are highly correlated. Instead, for explanatory models it becomes necessary to consider a theoretical causal model for how the various regressors and Y relate to one another. In particular, X_2 should be included in the model if it is a confounder, but not all variables highly correlated with the regressor of primary interest (i.e., X_1) are necessarily confounders. Vittinghoff et al. (2005), Jaccard, Guilamo-Ramos, Johansson, and Bouris (2006), and Hernan, Hernandez-Diaz, Werler, and Mitchell (2002) discussed various approaches for identifying whether a variable is a confounder and thus should be included in the regression model.

7. Sample Sizes Are Justified

Sample size is an important component to any research study. “Rules of thumb” that were once widely recommended for planning sample size are not generally appropriate and should not be used as justification (see Green, 1991, for a review). Instead, researchers should justify their sample size. A common approach to sample size planning is the power analytic perspective. However, another perspective is accuracy in parameter estimation (AICE). The goal of the power analytic approach is to plan sample size so that a false null hypothesis can be rejected with some desired probability

(i.e., power), whereas the goal of the AIPE approach is to obtain an accurate estimate of the population value, which is operationalized by a sufficiently narrow confidence interval with some desired degree of assurance (i.e., probability). In addition to deciding on whether power or AIPE is most appropriate, researchers also need to state whether the primary interest is in an omnibus effect (i.e., the squared multiple correlation coefficient) or one or more targeted effects (i.e., regression coefficients), which is necessarily based on the question(s) of interest. In particular, questions of prediction are more likely to involve omnibus effects, whereas questions of explanation are more likely to involve targeted effects. Additional details are provided in Kelley and Maxwell (2008), who discussed sample size planning methods in a multiple regression context in a 2×2 (power or AIPE \times omnibus or targeted effect) framework.

In some cases, existing/archival data become available to a researcher. Because the data have already been collected, sample size planning cannot be done as previously discussed, as it is implemented *a priori* in the design phase of the study. In general, power and AIPE are not often discussed for existing/archival data. However, power and AIPE can still be addressed, albeit in a different manner. In particular, for a specified value of an effect size at the size of the sample in the existing data, power and expected confidence interval width can be given. An appropriate value for the effect size to use is what can be termed the parameter of minimal importance (POMI) or the minimum parameter value of interest (MPVI), both of which represent the smallest magnitude that is deemed to have scientific, clinical managerial, or practical importance/interest in the particular context.

8. Missing Data

Missing data is a perplexing issue. There are three broad categories of missingness: (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR). MCAR is when missingness does not depend on either observed or missing values, whereas MAR is when missingness does not depend on the missing values but may depend on observed values. MNAR implies that missingness depends on an outside variable not in the model or depends on the value of the variable itself (see Little & Rubin, 2002, for a types of missing data and appropriate methods for dealing with the different types of missing data).

Although the specifics of the situation will differ, researchers should do their best to ensure the amount of missing data is minimized (e.g., remind participants about follow-up visits, check evaluations for blank responses before the participants leave, clearly state that sensitive data will remain confidential if appropriate). Generally, whenever missing data arises in a research study, it opens the possibility for criticism in the way it was (or was not) dealt with. Whenever there is a nontrivial amount of missingness, the data should be interrogated for patterns of missingness (Harrell, 2001). When apparent patterns are found, they should be reported and, if possible, a plausible explanation provided with a cautionary reminder given that exploratory methods were used to uncover any apparent patterns in the data. Regardless of the way in which missing data is dealt with, the method and the rationale for choosing the method should be discussed. That being said, some methods, in particular mean substitution and/or pairwise deletion, should not be used unless there is a good reason to do so with a clear explanation of why. We will briefly discuss three methods of dealing with missing data (see Schafer & Graham, 2002, for a thorough review).

When missing data does occur, casewise deletion appears to be frequently employed in the applied literature; however, casewise deletion can be problematic. In multiple regression, casewise deletion and listwise deletion are equivalent, however, in other models the two terms differ. Casewise deletion is when a participant is completely excluded, regardless of the amount of data available for the participant, if any data are missing for the analysis of a particular model. Listwise deletion is when an entire row is removed when there is any missing data. Thus, for models in which each case has

only one row in a data set, casewise and listwise deletion are equivalent. However, for some models a single case (e.g., person) will have multiple rows for different measurement occasions. There, such as in multilevel models, the row but not the case itself is deleted. Casewise (or listwise) deletion generally yields unbiased estimates only under the very strong assumption that data are MCAR. At best, estimates obtained using casewise deletion are inefficient, implying less statistical power and estimation accuracy than would otherwise be the case. The reason casewise deletion is inefficient is because the sample size is reduced to only those with complete data sets, which tends to increase the sample standard error(s) and necessarily does so in the population. More important, however, is that estimates obtained using casewise deletion will often be biased, unless plausible arguments can be advanced for why missingness is likely to be MCAR.

Imputation or multiple imputation provides a reasonable way to deal with missing data in many situations. Imputation is when a plausible value is substituted for a missing value and multiple imputation is when this process is performed multiple times. The “plausible values” come from an imputation model that uses other data that are available to estimate the data that are not available. At first the idea of estimating data might seem problematic, but it is often better to estimate what is usually a small amount of data than to disregard valuable data with deletion (e.g., casewise) strategies (Harrell, 2001, §3.4).

Full information maximum likelihood (FIML) and restricted maximum likelihood (REML) estimation are the most popular methods for dealing with missing data in multilevel models and structural equation models, likely because main-stream multilevel model and structural equation modeling programs can easily implement them (and usually do so by default). These maximum likelihood methods for dealing with missing data assume that data are MCAR or MAR. Because FIML does not consider the degrees of freedom and uses the standard normal distribution instead of the *t*-distribution, sample size should not be small with this approach. Small sample sizes being used with the FIML approach to missing data will tend to yield differences in the empirical and nominal Type I error rates. REML, however, does consider the issue of degrees of freedom and is more appropriate in smaller samples. Another issue is that maximum likelihood estimation assumes multivariate normality, which might not always be reasonable (recall that the standard multiple regression assumption is only that the errors are normally distributed). Enders (2001) provided a review and evaluation of maximum likelihood estimation when missing data exists in the context of multiple regression. Our recommendation is to use either multiple imputation or maximum likelihood estimation when faced with missing data.

9. Models Examining Moderation

The regression model shown in Equation (1) assumes that the effects of each X_k on Y are additive. For example, with two regressors, this model assumes that the relation between X_1 and Y is the same for every value of X_2 and similarly the relation between X_2 and Y is the same for every value of X_1 . In reality, however, the strength of the relation (or even the direction of the relation) between X_1 and Y might depend on X_2 , in which case X_1 and X_2 are said to *interact*. As a consequence, the regression model shown in Equation (1) might seem very restrictive, because it does not seem to allow for the possibility of an interaction between X_1 and X_2 . Fortunately, this restriction is illusory, because modifications to the model allow X_1 and X_2 to interact. The ability to modify this model is critical because many theories in the social and behavioral sciences stipulate that the relation between a pair of values (e.g., Y and X_1) depends on a third variable (e.g., X_2), which corresponds to an interaction effect.

The standard way of modifying the model in Equation (1) so as to allow for the possibility of an interaction (or equivalently, a moderator) is to add cross-product terms. For example, with two regressors, the model becomes

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}X_{2i} + \varepsilon_i \quad (4)$$

The inclusion of the product term allows the relation between either X and Y to depend on the value of the other X . In particular, this model stipulates that the slope relating X_1 to Y is given by

$$\frac{dY}{dX_1} = \beta_1 + \beta_3 X_2, \quad (5)$$

where dY/dX_1 is the derivative (instantaneous slope) of Y with respect to X_1 . If β_3 is non-zero, the relation between X_1 and Y depends on X_2 , so X_2 moderates the effect of X_1 on Y , or equivalently, X_1 and X_2 interact. However, many researchers might not realize that the product term represents a very specific type of interaction, namely a *bilinear* effect. In particular, Equation (5) shows that if β_3 is positive, the slope becomes increasingly higher for larger values of X_2 . Similarly, if β_3 is negative, the slope becomes increasingly lower for larger values of X_2 . Thus, researchers should consider whether this is the type of interaction they truly desire to detect. If not, more complicated models can be constructed, such as including quadratic terms for some or all regressors. Interested readers can consult Cohen et al. (2003) for additional details.

The best way to begin to interpret effects in moderator models is generally to plot the interaction. For example, suppose the primary interest involves the extent to which X_2 moderates the relation between X_1 and Y . Cohen et al. (2003) recommended plotting regression lines relating Y and X_1 at three values of X_2 (typically at the mean of X_2 and also at scores one standard deviation below the mean and one standard deviation above the mean). We recommend that such a plot be included in a published work involving moderator effects. Alternatively, what can be helpful is a three-dimensional representation of the relations, where Y is plotted as a function of all possible scores on X_1 and X_2 within an appropriate range.

A point of some confusion historically has been how to interpret the β_1 and β_2 coefficients in the model in Equation (4). Some researchers have interpreted these coefficients as if they corresponded to main effects, but this is not generally true. Instead, they are conditional (i.e., simple) effects. For example, Equation (5) shows that β_1 is the slope of Y on X_1 when X_2 equals 0. Unless the range of values of X_2 happens to include 0, the conditional effect in the interaction model will be meaningless. For this reason, it is often recommended that X_1 and X_2 be recoded so that a value of 0 takes on a meaningful interpretation. Most commonly, both variables are centered by subtracting the sample mean from all scores (*mean-centering*), yielding a new coding with a mean of 0. One could subtract a theoretically meaningful value from the scores. In any event, it is critical that authors explain how regressors in interaction models have been coded, in order to facilitate interpretation of the corresponding regression coefficients.

Because of perceived complications of interpreting interactions between continuous regressors, some researchers decide to simplify analyses by categorizing either or both regressors. We strongly recommend that researchers avoid the temptation to categorize continuous variables. One reason to leave variables as continuous is that categorization can decrease power. Interestingly, Maxwell and Delaney (1993) have also shown that in some situations categorization can have the opposite effect of producing spurious effects, thus inflating the Type I error rate. Thus, statistically significant interaction effects based on artificially categorized variables cannot necessarily be trusted, strengthening the argument for leaving continuous variables as continuous.

Researchers should also be aware that several other factors affect the ability to detect interactions in regression models. First, when X_1 and X_2 are measured with error, the product term X_1X_2 will generally be much less reliable than either X_1 or X_2 , which tends to lower the power to detect an interaction. Researchers who use regression to investigate interactions need to consider carefully

the reliability of regressors. Second, McClelland and Judd (1993) showed that the distribution of regressors in observational studies will often reduce power, especially when regressors correlate substantially with one another. Third, Lubinski and Humphreys (1990) showed that when regressors correlate substantially with one another, the Type I error for testing an interaction can be badly inflated if curvilinear effects exist but are not included in the regression model. Including higher order effects such as X_1^2 and X_2^2 can guard against spurious interaction effects, but also runs the risk of greatly lowering power to detect true interaction effects. There is no clear consensus among methodologists at this point about how best to resolve this dilemma. At the very least authors who want to investigate interactions in regression models should be clear about the extent to which their regressors correlate with one another as well as the extent to which theoretical considerations either do or do not rule out possible curvilinear effects. Given the scope of the topic of interactions, we recommend that readers consult such sources as Aiken and West (1991) and Jaccard and Turrissi (2003) for further information, as well as Chapter 18 in this volume.

10. Models Examining Mediation

Baron and Kenny (1986) clarified the distinction between moderation and mediation. Both involve a role that X_2 (for example) may play in the relation between X_1 and Y , leading some researchers to confuse moderation and mediation. Thus, it is incumbent on authors of papers reporting either moderation or mediation to provide a clear theoretical rationale for their study.

The variable X_2 mediates the relation between X_1 and Y when X_1 causes X_2 and X_2 in turn causes Y . Thus, mediation can be represented by a pair of regression models:

$$X_{2i} = \beta_0^* + \beta_1^* X_{1i} + \varepsilon_i^* \quad (6)$$

$$Y_i = \beta_0 + \beta_2 X_{1i} + \beta_3 X_{2i} + \varepsilon_i, \quad (7)$$

where the asterisk represents values from the model where X_2 is the dependent variable with X_1 as its regressor. From this perspective, X_2 is a mediator when both β_1 and β_3 are non-zero. In the special case where β_2 equals 0, X_2 is said to completely (or fully) mediate the relation between X_1 and Y ; otherwise, X_2 partially mediates the relation.

Baron and Kenny (1986) suggested a four-step procedure for establishing mediation. Subsequent research has studied their approach as well as a variety of alternatives. This is an area of continuing methodological research, and at this point either of two different approaches seems advisable for establishing mediation. One approach involves bootstrap methods (Shrout & Bolger, 2002). The other involves the distribution of the product variable $\beta_1\beta_3$ (MacKinnon, Lockwood, & Williams, 2004). We recommend that authors use either of these two methods to test mediation. Authors should also report coefficients and corresponding confidence intervals for relevant parameters as shown in Equations (6) and (7).

Several other factors should be considered in a mediation analysis. First, it is well known that error of measurement in the mediator causes biased estimates of regression coefficients. In three-variable models such as those in Equations (6) and (7), random measurement error will tend to result in an underestimate of the mediated effect and an overestimate of the direct effect of X_1 on Y . Researchers should address this likely bias in any interpretation of their results unless the mediator is measured without error. Alternatively, a latent variable model might be used in order to address measurement error and its biasing effects. Second, Maxwell and Cole (2007) have shown that cross-sectional estimates of mediation can be seriously biased when mediation occurs over time. Researchers who rely on cross-sectional analyses need to interpret their results with appropriate caution, and should be

encouraged to consider longitudinal designs instead of cross-sectional designs. Third, researchers should carefully consider necessary sample size to obtain adequate power. Fritz and MacKinnon (2007) provided useful guidelines. Fourth, further information about mediation, especially for more complicated models with more than three variables, is available in Chapter 18 of this volume and in MacKinnon, Fairchild, and Fritz (2007) and MacKinnon (2008).

11. Checking Assumptions Visually

The assumptions of the multiple regression model should be considered and evaluated whenever the model is used. As Anscombe (1973) noted, graphs can help researchers appreciate broad features of data *and* look beyond broad features to literally see potentially unexpected relationships, outliers, and violations of assumptions, et cetera. Anscombe went on to show four very different figures, three of which have gross violations of multiple regression assumptions, yet where the results from the regression model were the same (i.e., estimates, p -values, confidence intervals, etc.). Recall that the linearity assumption is that the expected value of Y given the K regressors is a linear function of the K variables. We recommend a *conditioning plot* (also referred to as a *coplot*) for examining the critical assumption of linearity. Another useful set of plots for this purpose are *residual versus predictor* (RVP) and *component plus residual* (CPR) plots. One way to evaluate violations of this assumption for “obvious” violations is by plotting the residuals as a function of the model implied values. An obvious nonlinear relationship is evidence that the linearity assumption does not likely hold. When such is the case, there might be an important variable not included in the model, an interaction term might be appropriate, or the relation between the K regressors and the criterion might be nonlinear in nature. As previously noted, the latter, in our opinion, is not considered frequently enough, and correspondingly nonlinear models are not applied in many areas as often as we believe that they should be, based on theory and empirical evidence. For example, sigmoidal forms or asymptotic values cannot adequately be modeled with linear models. We suggest readers consult Seber and Wild (1989) for a discussion of nonlinear regression models.

Recall that the errors in a multiple regression model fitted with ordinary least squares are assumed to be normally distributed for the validity of the significance test and confidence intervals. A normal-quantile–empirical-quantile plot (generally termed a *qq-plot*) is a two-dimensional plot where theoretical quantiles from the normal distribution are compared to the empirical quantiles of the observed errors. The qq-plot allows a visual evaluation of the assumption of normality of the errors. Gross violations of the normality assumption of the errors can often easily be seen with the use of a qq-plot. Although there are formal statistical tests to evaluate normality, visual displays are often extremely effective at identifying potential problems and are often easier to implement and interpret.

Matrix scatterplots (sometimes called *pairs plots*) are helpful to examine the bivariate relations among the $K + 1$ variables. These plots can also reveal observations that might be miscoded or identify potential outliers. Further, those cases that might not be considered outliers on either of two variables individually might be an outlier in a bivariate sense (which could heavily influence estimation and inference). For example, if there is a strong positive relation between X_1 and Y , yet one observation has a very low X_1 value and a very high Y value, that point would disproportionately affect the estimate of the line of best fit (e.g., Cohen et al., 2003, for a review). Such a case would not be readily identified without visualization (or more formal outlier/influential data point checks), which could allow the possibility of further investigating such a unique case. Cases in such situations are said to be *leveraging points*. In general, formally operationalizing what constitutes an outlier and appropriately dealing with them can be difficult, but it is nevertheless important. Cohen et al. (2003, ch. 10) provided a detailed discussion of possible causes and possible remediations when outliers are believed to exist. Regardless of the exact way in which outliers are dealt with, transparency to the reader is key. Transparency is

especially important because two researchers analyzing the same data might come to different conclusions when fitting the same model based only on how outliers are addressed.

In published work, space is often at a premium, which has the effect of only infrequently printing figures that evaluate the model assumptions (e.g., RVP, CPR, qq-plots). Nevertheless, even if such figures are not part of the published version of a work, there is little question that they can be very beneficial for authors, as well as satisfying reviewer curiosity on model fit and appropriateness, and can help to convey relationships to the reader easier seen than said. We think it is generally wise for authors to include a brief discussion of the (published or unpublished) figures and the seemingly appropriateness or inappropriateness of the model. Of course, if the figures help to identify weakness in the appropriateness of the model, other models should be considered and such a finding noted in the work. In short, visualization techniques should help justify the model chosen and this information should be conveyed to readers.

We are sensitive to the amount of journal space that such plots can consume. Due to limited journal space, editors may be reluctant to allow several pages of figures, even if they are informative. We believe a reasonable solution is for authors to produce supplemental material that can be referenced in the article but stored on a journal's supplemental materials web page, which many journals now make available. If not on a journal supplements page, the author(s) can often post additional information on an archival site (e.g., via university library).

12. Measurement Error

Measurement error in multiple regression can be conceptualized in a $2 \times 2 \times 2$ array, where depending on the specific conditions the effect of measurement error has different implications. The dimensions of the array are (a) type of measurement error (random or nonrandom), (b) type of variable (regressor or criterion), and (c) type of coefficient (unstandardized or standardized). We will briefly describe each dimension of the array below.

Random measurement error, which is omnipresent in research, is uncontrolled error that is assumed to have a mean of zero. Nonrandom measurement error, however, work will tend to have a mean that is not zero and/or be correlated with errors. In short, nonrandom measurement error in the criterion and/or the regressor is problematic and can lead to biased estimates of model parameters. Because nonrandom measurement errors often represent a flaw in the measurement procedure, instrument, or design, we will simple say that multiple regression is not generally appropriate in circumstances of nonrandom measurement error, with the exception being when the nonrandom error is so small that is has essentially no effect on the mean and covariance structures of the variables.

We will assume the random measurement errors have a mean of zero and are uncorrelated with measured variables, with their corresponding true scores, and with all other errors. Provided the regressors are unstandardized, any measurement error in Y is absorbed into the model error term, from Equation (1), and has no effect on the expected value of the regression coefficients. Thus, under the standard multiple regression assumptions, the regression coefficients remain unbiased. However, because the model error variance increases, the estimate of the squared multiple correlation coefficient is systematically lowered. Because R^2 decreases—it is attenuated due to a larger error variance—the standard errors of the regression coefficients will also be larger, implying that statistical power and the accuracy of parameter estimates are reduced via a decrease in precision. However, in the situation where the regression model is standardized, the regression coefficients will be attenuated when the criterion is measured with error (Kenny, 1979). The attenuation occurs when the criterion is measured with error because for standardized regression coefficients the multiplier (i.e., s_{xk} / s_Y for the k th regressor) of the unstandardized regression coefficient that yields the standardized regression coefficient has a denominator whose expected value is larger than the true

value. The expected value of s_Y is larger than σ_{Y_r} , the population standard deviation of the true scores of Y . From a classical test theory perspective on random measurement errors, the variance of Y is the sum of the true score variance ($\sigma_{Y_r}^2$) and the error variance ($\sigma_{Y_e}^2$). Thus, will tend to be larger than σ_{Y_r} , which leads to observed standardized regression coefficients smaller than their corresponding true values (Kenny, 1979, ch. 5).

In observational research, the case of random measurement error in one or more regressors will generally lead to biased regression coefficients, regardless of whether or not the regressors are standardized. As Fox (2008) showed, in simple regression (i.e., when $K = 1$) when measurement error occurs in the (only) regressor, its regression coefficient is generally attenuated. However, with one exception, no general statement can be given for the effect of measurement error in one regressor on the regression coefficient for the other regressors in a multiple regression model (i.e., when $K > 1$). As Kenny (1979, p. 104) pointed out, measurement error in one regressor can attenuate regression coefficients, make the estimate of a regression coefficient that is zero be nonzero, and can change the sign of a regression coefficient. The exception noted is for designed experiments, where the randomly assigned variable is uncorrelated with other regressors in the model. When the randomly assigned variable has measurement error, the regression coefficient is less accurate; it is unbiased but less precise. Because the regression coefficient is less precise, the corresponding confidence interval tends to be wider and the test of the null hypothesis will not be as powerful (larger p -value).

In general, the difficulty in saying what happens when measurement error occurs in an observational application of multiple regression lies in the multivariate nature of multiple regression, as the properties of one regressor influence the regression coefficients of all other regressors. In short, when a regressor is measured with error in an observational application, its effects are not partialled out as fully as when it is measured without error. This concept is easiest to understand when one regressor is perfectly unreliable, and thus the effects of the true regressor have not been partialled in any way (Kenny, 1979). As a result, the coefficients for other regressors in the model are generally biased because the perfectly unreliable regressor has not been controlled for at all. The important point is that whenever a regressor is measured with error, not only is the coefficient associated with that regressor biased, but typically so are all of the other coefficients in the model, including even coefficients for any regressors that happen to be measured without error. Because the value of the regression coefficient for the variable that is measured with error is biased, being smaller in magnitude than it otherwise would have been if the variable were perfectly reliable, the bias will generally lead to an error variance larger than it would have been, which then leads to a negatively biased estimate of P^2 (i.e., R^2 is, on average, smaller than it should be), ultimately leading to larger standard errors for all of the regression coefficients in the model.

It is desirable to minimize measurement error in all uses of multiple regression. However, measurement error is especially problematic when the primary goal is explanation, because theoretical explanations virtually always relate to constructs, not to variables measured with error. When confronted with nontrivial measurement error, it is often advisable to obtain multiple measures of each construct and use structural equation modeling (see Chapter 33, this volume) instead of multiple regression. Measurement error can be less problematic when the goal is prediction, because the practical goal is often to determine how well regressors as measured can predict the criterion as measured. When the goal is explanation and nontrivial measurement error is likely to occur, we generally recommend obtaining multiple measures of each construct so that structural equation modeling can be used.

13. Statement of Limitations

Multiple regression is a flexible system for linking K regressor variables to a criterion variable of interest. In many cases, multiple regression is an appropriate statistical model for addressing common research

questions, whether they be for purposes of explanation, prediction, or both. Nevertheless, multiple regression has limitations that are defined in part by the model and its assumptions as well as by the research design. The limitations of multiple regression in the specific context should be discussed.

Multiple regression has limitations, like other statistical models, when attempting to infer causality from a research design that was not experimental in nature (i.e., when random assignment of levels of the regressors to the participants was not part of the design). Although including additional regressors that are thought to be correlated with the regressor of interest adds a form of statistical control, with regard to causality there is no way to “control” all possible confounders unless randomization is an explicit part of the design. In purely observational designs, claims of causality should generally be avoided. The benefits of randomization cannot be overemphasized, even if for only some of the variables in the design, because randomization implies that the participants have equal population properties (e.g., mean and covariance structures) on all outside variables.

Variables termed “control” variables are often included in multiple regression, as previously noted. However, including a control variable in the model in no way implies that the variable can literally be “controlled”—use of such a term is based on a precise statistical meaning and is not literal in the sense of everyday language. When something is “controlled for” it allows for the linear effect of each regressor on the criterion variable to be evaluated (i.e., a regression coefficient estimated), while holding constant the value of the other regressor variables. In practice, however, many variables cannot be controlled by the researcher, even in the most carefully designed studies. Thus, there is not literally any control by the researcher in an observational design over the variables said to be “controlled for.” Rather, an effect can be examined while holding constant the other variables.

The reasonableness of temporal ordering of variables needs to be considered, as multiple regression can be applied in ways such that an explanatory variable is nonsensically used to model a criterion variable. Although the multiple regression model may account for a large proportion of variance, it might not make theoretical sense. For example, multiple regression could be used to model “time spent studying” as a function of “test score.” However, such a model is nonsensical in the sense that “time spent studying” would be an explanatory variable of “test score.” This is a simple example of a causality problem, in the sense that the multiple regression model itself does not make a distinction between what causes what. Theory, of course, should be the guiding principle of the specification and direction of causal relationships. Inferring causality can be difficult, especially because there technically needs to be some passage of time that occurs in order for a regressor to literally cause some change in a criterion (unless simultaneous causality is presumed).

14. Alternatives to Multiple Regression

When the assumption of normality of errors is violated, nonparametric approaches to inference for multiple regression should be considered (e.g., Efron & Tibshirani, 1993; Györfi, Kohler, Krzyzak, & Walk, 2002). Multiple regression assumes that outcome variables are continuous and observed. However, when the criterion variable is censored, truncated, binary/dichotomous, ordinal, nominal, or count, an extension of the general linear model termed the generalized linear model, where a link function (e.g., exponential, Poisson, binomial, logit) relates the linear regression equation (analogous to the right hand side of Equation (1)) to a function of the criterion variable (e.g., probability of an affirmative response) can be used (e.g., Agresti, 2002; Long, 1997; McCullagh & Nelder, 1989; Chapters 16 and 17 in this volume).

Linearity is an assumption that is not reasonable in some situations, either based on theoretical or empirical evidence (e.g., the graphical displays previously discussed). *Spline* regression models allow different slopes over ranges of one or more regressors, in what has appropriately been termed a piecewise model (e.g., Fox, 2000; Ruppert, Wand, & Carroll, 2003). In spline regression multiple

“knots” exists, where the slope of the regression line (potentially) changes over specified ranges (note that the slopes can be discontinuous in that they need not overlap at a knot). Another nonparametric regression procedure is known as *lowess* (locally weighted scatterplot smoothing) (also denoted *loess*; e.g., Cleveland, 1979; Fox, 2008), in which multiple regression models are fitted to areas/regions of the regressor(s) with “local” points receiving more weight than more distant points. The definition of “local” changes as a function of the width of the span selected, which is a parameter in the control of the analyst and for which there is not a single best answer to the ideal size of the span. For short spans the line of best fit can differ dramatically over a small range of a predictor, whereas a wide span tends to have a relatively smooth relationship between the regressor(s) and the criterion. Lowess techniques are most often used when $K = 1$. More general than lowess models are generalized additive models that allow some regressors to enter the model linearly and some to enter as splines (Ruppert et al., 2003, p. 215).

Applications of the general linear model are not robust to violations of the assumption of independent observations. Even for the simple case of the two independent group *t*-test, which can be considered a special case of multiple regression, it is known that the nominal and empirical Type I error rate can be drastically different when the assumption of independence is violated (e.g., Lissitz & Chardos, 1975). When observations are not independent (e.g., students nested within classrooms, clients nested within therapists, observations nested within person), appropriate methods to explicitly control for the lack of independence should be used. A general approach to handling such nonindependence is multilevel models (also termed *hierarchical linear models*, *mixed effects models*, or *random coefficient models*; see Chapter 22, this volume).

When measurement error is not ignorable, multiple regression is not ideal and latent variable models should be considered, especially when the primary goal is explanation instead of prediction. In particular, confirmatory factor analysis (see Chapter 8, this volume) and structural equation modeling (see Chapter 33, this volume) allow for explicitly incorporating error into the model of interest, which has the effect of separating the “true” part of the model from the “error” part.

References

- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Algina, J., & Olejnik, S. (2000). Determining sample size for accurate estimation of the squared multiple correlation coefficient. *Multivariate Behavioral Research*, 35, 119–136.
- Ancombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27, 7–21.
- Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, 8, 129–148.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Burket, G. R. (1964). *A study of reduced rank models for multiple prediction*. Psychometric Monograph, no. 12. Richmond, VA: Psychometric Corporation. Retrieved from www.psychometrika.org/journal/online/MN12.pdf.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, 69, 161–182.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- Enders, C. K. (2001). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, 61, 713–740.
- Ezekiel, M. (1930). *Methods of correlational analysis*. New York: Wiley.
- Fox, J. (2000). *Multiple and generalized nonparametric regression* (No. 131). Thousand Oaks, CA: Sage.
- Fox, J. (2008). *Applied regression analysis, linear models, and related methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18, 233–239.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

- Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, 26, 499–510.
- Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2002). *A distribution-free theory of nonparametric regression*. New York: Springer.
- Harrell, Jr., F. E. (2001). *Regression modeling strategies with applications to linear models, logistic regression, and survival analysis*. New York: Springer.
- Hernan, M. A., Hernandez-Diaz, S., Werler, M. M., & Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology*, 155, 176–184.
- Huberty, C. J. (2003). Multiple correlation versus multiple regression. *Educational and Psychological Measurement*, 63, 271–278.
- Jaccard, J., Guilamo-Ramos, V., Johansson, M., & Bouris, A. (2006). Multiple regression analyses in clinical child and adolescent psychology. *Journal of Clinical Child and Adolescent Psychology*, 35, 446–479.
- Jaccard, J., & Turrissi, R. (2003). *Interaction effects in multiple regression* (2nd ed.). Thousand Oaks, CA: Sage.
- Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, 20, 1–24.
- Kelley, K., & Maxwell, S. E. (2008). Power and accuracy for omnibus and targeted effects: Issues of sample size planning with applications to multiple regression. In P. Alasuuta, J. Brannen, & L. Bickman (Eds.), *Handbook of social research methods* (pp. 166–192). Newbury Park, CA: Sage.
- Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley-Interscience. Lissitz, R. W., & Chardos, S. (1975). A study of the effect of the violation of the assumption of independent sampling upon the Type I error rate of the two-group t-test. *Educational and Psychological Measurement*, 35, 353–359.
- Little, R. J. A., & Rubin, D. A. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley and Sons.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Lubinski, D., & Humphreys, L. G. (1990). Assessing spurious “moderator effects”: Illustrated substantively with the hypothesized (“synergistic”) relation between spatial and mathematical ability. *Psychological Bulletin*, 107, 385–393.
- MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7, 19–40.
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Erlbaum.
- MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, 58, 593–614.
- MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39, 99–128.
- Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, 12, 23–44.
- Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, 113, 181–190.
- Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). New York: Routledge.
- McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114, 376–390.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman & Hall.
- Muller, K. E., & Fetterman, B. A. (2002). *Regression and ANOVA: An integrated approach using SAS Software*. Cary, NC: SAS Institute.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Orlando, FL: Harcourt Brace.
- Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement*, 23, 99–115.
- Rencher, A. C., & Pun, F. C. (1980). Inflation of R^2 in best subset regression. *Technometrics*, 22, 49–54.
- Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics* (2nd ed.). Hoboken, NJ: Wiley.
- Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. New York: Cambridge University Press.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177. Seber, G. A. F., & Wild, C. J. (1989). *Nonlinear regression*. New York: Wiley.
- Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, 7, 422–445.
- Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9, 164–182.
- Steiger, J. H., & Fouladi, R. T. (1992). R2: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. *Behavior Research Methods, Instruments, and Computers*, 4, 581–582.
- Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2005). *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models*. New York: Springer.
- Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.

24

Multitrait–Multimethod Analysis

Keith F. Widaman

Campbell and Fiske (1959) argued that every measurement we obtain in psychology is a trait-method composite—a measure purportedly of a particular trait construct obtained using a given method of measurement. Campbell and Fiske introduced the multitrait–multimethod (MTMM) matrix as a tool for evaluating systematically the correlations among a set of measures obtained using multiple methods. The primary utility of the MTMM matrix approach is the opportunity such a study affords to determine the preponderance of trait-related and method-related variance in measures in a battery. To aid in this evaluation, Campbell and Fiske argued that researchers should measure each of t traits (e.g., Extraversion, Neuroticism, Fluid Intelligence) using each of m methods (e.g., self-report, objective tests, observer ratings), so that each trait is measured using each method. By arranging trait measures in the same order within methods, the MTMM matrix should exhibit clear patterns to satisfy the dictates of convergent and discriminant validation. Convergent validation is satisfied if the researcher finds high correlations among measures of putatively the same construct using different methods of measurement, and discriminant validation is satisfied if low correlations are found among measures of presumably different constructs. Campbell and Fiske described several rules of thumb for evaluating patterns of correlations in the MTMM matrix. Specifically, (a) correlations between measures of the same construct obtained using different methods of measurement should be large; (b) correlations between measures of the same construct obtained using different methods of measurement should be larger than correlations of those measures with measures of different constructs obtained using the same or different methods; and (c) the same pattern of trait correlations should hold for all combinations of methods.

Among others, Jöreskog (1971) pioneered the fitting of confirmatory factor analysis (CFA) models to MTMM data. The CFA approach circumvented several problems associated with the Campbell and Fiske (1959) rules of thumb. In particular, the CFA approach (a) yielded clear significance tests of differences between alternative models and of specific parameter estimates, whereas the ordinal comparisons involved in the Campbell-Fiske rules of thumb relied on dependent comparisons that compromised statistical tests; (b) allowed for tests of the amount of trait-related and method-related variance in the MTMM matrix; and (c) led to estimates of the amount of trait-related and method-related variance in each measure. Widaman (1985) systematized earlier work on CFA models and provided an informative taxonomy of models for MTMM data by cross-classifying available trait factor structures and method factor structures. In addition, Widaman discussed alternate analytic strategies

for exploring the magnitude of effects of trait and method constructs underlying manifest variables in an MTMM matrix. Building on earlier work by Kenny (1976), Marsh (1989) discussed an additional CFA method specification, using correlated uniquenesses to represent method effects. At about this time, Browne (1984) described a multiplicative model for fitting MTMM data. Reichardt and Coleman (1995) provided an advanced discussion of the relative fit of linear and multiplicative models. Eid and colleagues (Eid, Lischetzke, Nussbeck, & Trierweiler, 2003) discussed an approach leaving out one method factor to improve identification of model parameters. Recent advances to improve the identification and estimation of CFA models for MTMM data have been made by Castro-Schilo, Grimm, and Widaman (2016) and by Helm, Castro-Schilo, and Oravecz (2016), which should be consulted for the most state-of-the-art approaches and issues regarding model specification and estimation.

Table 24.1 Desiderata for MTMM Analysis.

<i>Desiderata</i>	<i>Manuscript Section(s)*</i>
1. Identify and justify nature of trait constructs under consideration.	I
2. Discuss the ways in which the methods of measuring traits might influence measurements.	I
3. Discuss the usefulness of trait and method variance estimates for understanding the underlying nature of measurements included in study.	I
4. Present path diagrams for alternative models to be considered or for a general model within which alternative models are nested.	I
5. Describe the general strategy that is followed when comparing competing models.	I
6. Describe nature and size of participant sample, along with basic descriptive data.	M
7. Describe the number of items, response scale, etc. for each manifest variable.	M
8. Note the manifest variables associated with each trait and method factor, verifying that each factor has a sufficient number of indicators and note method of identification.	M
9. Give name and version of program used, state the method of estimation used, and address assumptions.	M
10. Describe the origins of missing data or outlier values on manifest variables and how these are handled.	M, R
11. Describe how problems (lack of convergence, improper estimates) are handled.	M, R
12. Provide the MTMM matrix, along with means and standard deviations of manifest variables.	R
13. Follow an analytic strategy developed for the empirical application, which may dictate that certain (not all) possible models are evaluated.	R
14. Use likelihood ratio chi-square, practical fit indices, justify any needed “fixes” to circumvent problems in the model fitting.	R
15. Describe and justify theoretically and statistically any needed post-hoc modifications.	R
16. Evaluate parameter estimates for statistical significance and with regard to interval estimate (i.e., consider standard errors for each estimate).	R
17. Compute the proportion of variance due to trait, method, and unique factors.	R
18. Discuss the quality of the different manifest variables for representing trait and/or method influence.	D
19. Discuss how current results might impact future research on the traits and the substantive domain.	D
20. Discuss any limitations of the current study.	D

* Note: I = Introduction, M = Method, R = Results, D = Discussion.

Lance, Noble, and Scullen (2002) and Eid and Diener (2006) provided overviews of the strengths and weakness of different CFA models. Stacy, Widaman, Hays, and DiMatteo (1985) and Widaman, Stacy, and Borthwick-Duffy (1993) conducted empirical studies with non-standard method factor specifications. Specific desiderata for studies that utilize CFA models for MTMM data are presented in Table 24.1 and then discussed. This chapter deals only with the structural equation modeling approach to evaluating MTMM matrices. Readers interested in other approaches can consult other references for more traditional or historical approaches. For the original approach involving comparing correlations, readers should refer to Campbell and Fiske (1959) or Ostrom (1969). Hubert and Baker (1978) presented a nonparametric approach for testing the patterns of differences among correlations in an MTMM matrix that improved statistically on prior methods. Analysis of variance (ANOVA) approaches have also been proposed, and the relative strengths and weaknesses of ANOVA relative to CFA modeling were discussed by Millsap (1995). An overview of alternative methods for analyzing MTMM data was provided by Eid (2006).

1. Substantive Context and Measurement Implications

One initial and important goal of any investigation that uses the MTMM matrix is to identify and justify carefully the nature of the trait constructs that are the focus of the study. Trait constructs employed in psychology come in many different forms. In the mental ability domain, constructs such as fluid and crystallized intelligence or spatial ability can be studied, and researchers typically assume that these dimensions represent characteristics of persons that are rather stable across time. In the personality domain, researchers have recently emphasized the “big 5” constructs of extroversion, agreeableness, conscientiousness, neuroticism, and openness, which are also hypothesized to be stable characteristics. However, theory associated with certain other personality dimensions, such as state anxiety, presumes that individual differences on these dimensions will exhibit notable fluctuation across time. In still other areas of study, the trait constructs may represent internal psychological processes, behavioral interaction styles, and so forth. Regardless of the constructs under investigation, the nature and definition of the constructs must be considered carefully, especially whether assessing each construct with each method of measurement is appropriate and theoretically justified.

Most state-of-the-art approaches to analyzing MTMM data involve confirmatory factor analysis (CFA; see Chapter 8) or other sophisticated approaches to analyses, which fall within the general family of methods of structural equation modeling (SEM; see Chapter 33). As will be discussed later (see Desideratum 9), the most common method of estimating parameters in CFA or SEM is maximum likelihood estimation, which requires that data satisfy certain assumptions. As a result, the researcher should verify that the measures of the trait constructs under study have been previously subjected to exploratory or confirmatory factor analyses or other forms of psychometric analysis to verify their basic psychometric properties. Chief among these properties are normal distributions of scores on manifest variables and linearity and bivariate normality of relations among manifest variables. If data fail to satisfy these assumptions, other methods of estimation (e.g., robust weighted least squares) can be used that require less stringent assumptions. In addition, as discussed below (see Desideratum 9), the study should include at least three trait constructs in the MTMM matrix, to ensure adequate identification of latent variables.

2. Methods of Measurement and Their Potential Impacts

To conduct an MTMM study, the investigator must select a set of methods, preferably three or more, to ensure identification of the method latent variables (see Desideratum 9). The researcher should also consider and discuss the ways in which the methods of measurement included in the

study might influence those measurements of individual differences on trait dimensions. This is a difficult task, as a general theory of method effects, specifically in the context of MTMM studies, has never been developed. Despite the lack of a general theory of method effects, researchers have used many different ways of operationalizing measurement methods, and the brief summary below is offered to assist researchers in the challenging but essential task of understanding and interpreting method effects.

Many different types of measurement methods have been used in MTMM studies. In early studies of multiple dimensions of attitude (e.g., Ostrom, 1969), the methods included different ways of formatting items and/or developing scales (e.g., Guttman scaling, Thurstone scaling, rating scales). Some studies utilize different reporters as representing the different methods, and these can take any of several forms: (a) self, parent, and teacher report methods; (b) self, friend, and observer report methods; (c) mother, father, and teacher report; and so on. Still other studies survey methods more broadly, using life history (L), observer ratings (O), self-report (S), and objective test (T) indicators to serve as methods. These LOST indicators clearly represent a much broader selection of methods, as most researchers select methods within a single LOST category to serve as the multiple methods of measurement.

Methods may have many impacts on the trait measurements, an underemphasized aspect of most studies using the MTMM matrix. Self-reports may be contaminated by notable amounts of response biases such as acquiescence, social desirability, or extremity bias. Observer reports are likely to reflect halo bias, or the tendency to rate a particular target similarly across trait ratings, particularly if the trait constructs have close connections (e.g., effort and quality of performance on the job). Objective tests are highly influenced by motivation to perform well at the time of the assessment, and such motivation may wax and wane over time. Life history data are more indicative of typical levels of performance than of maximal performance, which may lead to lower levels of convergence with objective test scores. The list of potential influences of methods on measurements is large, and research studies will be improved in the future if they are designed to shed light on the alternative sources of variance in method effects.

3. Utility of Trait and Method Variance Estimates

Any findings regarding the trait and method variance in each of a set of measures is, of itself, an important contribution to the literature, especially if such estimates are not widely available from prior research. High levels of trait-related variance in measures support conclusions that the measures reflect the processes or constructs hypothesized, whereas low levels of trait-related variance should signal the need to revise measures to capture more adequately the underlying constructs. Either way, the use of the MTMM matrix can provide crucial information for interpretation of research beyond the scope of the current study.

The researcher should discuss the usefulness of trait and method variance estimates for understanding the nature of measurements included in study. The history of every area of psychology is a history strewn with examples of theories built upon measures that were presumably reflective of specified underlying processes or theoretical constructs, measures later shown to be only weakly related to the underlying processes hypothesized. Much wasted effort might have been avoided if researchers had utilized MTMM matrix studies to investigate the properties of their measures.

Most research in psychology has a built-in confirmation bias, as researchers tend to search out and highlight positive correlations among measures of similar constructs. Some of these positive correlations are statistically significant, but fall in the range of .30 to .40, and correlations of this magnitude are not strong evidence that the measures are indicators of the same construct. Furthermore, finding low correlations between disparate measures can be just as important as, or

more important than, finding high correlations between similar constructs. If a researcher found a .40 correlation between two measures of Construct 1, but found that both of these measures also correlate .40 with a measure of Construct 2, the researcher should be wary of the strength and importance of the former correlation. Use of the MTMM matrix approach forces researchers to confront research outcomes of this sort.

4. MTMM Path Diagrams

A path diagram or structural modeling diagram is a graphical presentation of the form of a statistical model. Most common linear models can be formulated as diagrams, and the statistical model represented by the diagram is often isomorphic with the diagram. In a path diagram, we typically denote manifest or measured variables as squares or rectangles and latent (or unmeasured) variables as circles or ellipses. Straight, single-headed arrows are used to indicate dependence or directed relations between variables, with the variable at the tail of the arrow a predictor of the variable at the head of the arrow. Finally, double-headed (and often curved) arrows are used to denote undirected relations, such as the covariance between two variables or the variance of a variable. In particular, a curved, doubled-headed arrow from a variable to itself represents a residual variance, with effects of all unidirectional influences on the variable controlled statistically. For example, in Figure 24.1, the curved, double-headed arrow from Trait Factor 1 to itself represents the entire variance of this trait factor because no unidirectional arrows are drawn toward the trait factor. In contrast, the curved, double-headed arrow from the manifest variable “Trait 1 Method A” to itself represents the unique variance of this indicator, which reflects variance in the indicator remaining after accounting for variance due to Trait Factor 1 and Method Factor A.

In a study using the MTMM matrix, the researcher can and usually should present one or more path diagrams for alternative structural models to be investigated in the study. Two alternative path diagrams are shown in Figures 24.1 and 24.2. In Figure 24.1, nine manifest variables are shown in the rectangles in the middle of the figure: Traits 1, 2, and 3 each assessed using Methods A, B, and C. In the standard correlated trait–correlated method (CTCM) model, the researcher can hypothesize the presence of three trait latent variables, each associated with the manifest variables aligned with the trait construct and shown in the circles on the left side of the figure. Thus, Trait Factor 1 is shown having direct linear relations on the three manifest indicators of Trait 1, Trait Factor 2 has direct linear relations on the three manifest indicators of Trait 2, and so forth. The potential influence of three method factors is also represented by the three ellipses at the right side of Figure 24.1. Method Factor A is presumed to have direct linear effects on all manifest variables measured using Method A, and similar direct relations hold for Method Factors B and C. The double-headed arrows among the three trait factors reflect covariances (or correlations) among these latent trait factors, and the double-headed arrows among the three method factors reflect covariances among the latent method factors. The absence of double-headed arrows between trait and method factors indicates that these covariances are presumed to be nil and are typically forced to be zero. We note that the model proposed by Eid (2000), which has become known as the correlated trait–correlated method minus 1, or CTC(M – 1), model and is recommended by some, can be obtained simply by deleting one method factor from the model shown in Figure 24.1.

The model in Figure 24.2 is often termed the correlated trait–correlated uniqueness (CTCU) model. As with the first model, nine manifest variables are shown in the rectangles in the figure. The trait factor specification in Figure 24.2 is identical to that in Figure 24.1, with trait factors having direct effects on their respective manifest variables and covariances posited among the trait factors. The major difference between the two figures is that method factors have been deleted in Figure 24.2 and have been replaced by covariances among unique factors or uniquenesses. That is, covariances

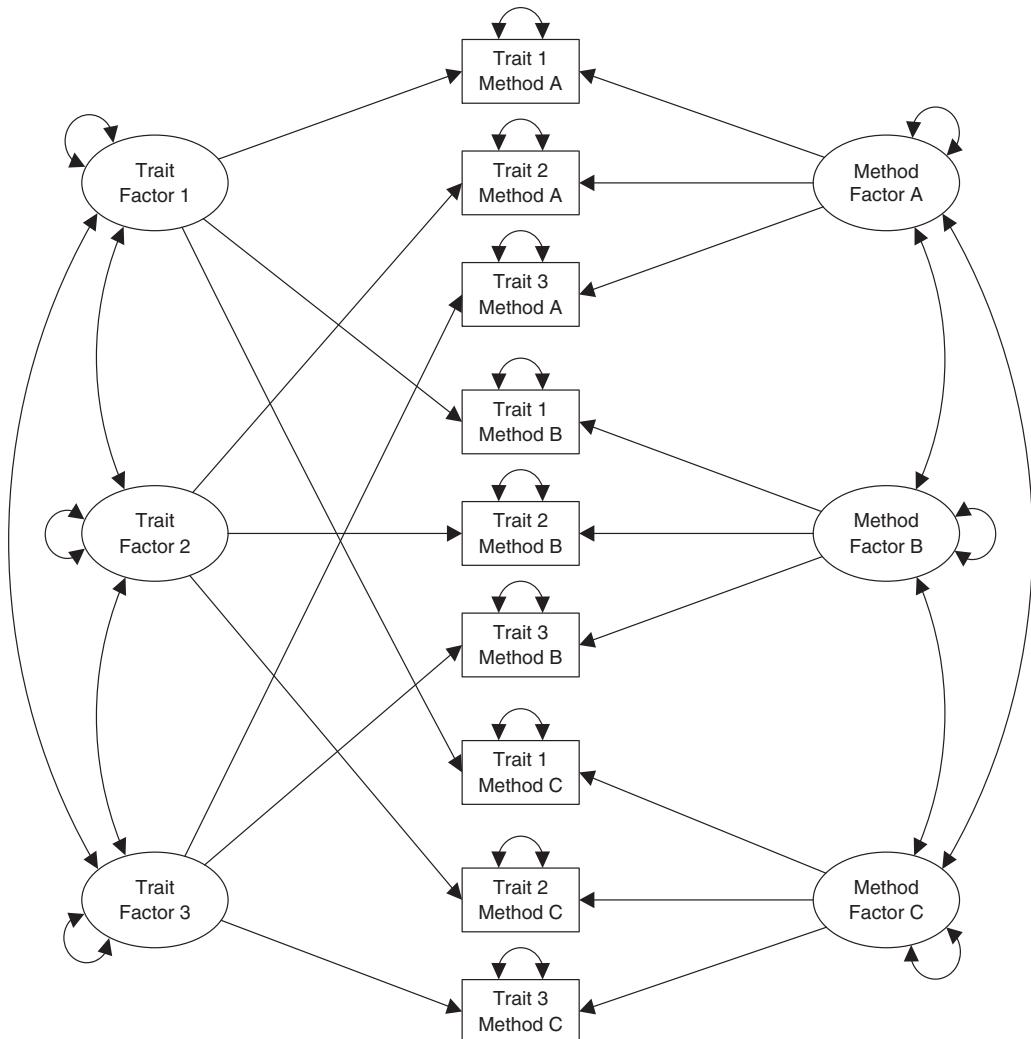


Figure 24.1 The Correlated Trait–Correlated Method (CTCM) Linear CFA Model for MTMM Data.

have been specified among all measures obtained using Method A, among all measures obtained using Method B, and among all measures obtained using Method C. Note that the absence of double-headed arrows between measures obtained from different methods (e.g., no double-headed arrows between measures under Method A with measures under Method B) means that method effects are hypothesized to be statistically uncorrelated under the CTCU model and such effects are therefore fixed at zero.

5. Analytic Strategy for Comparing Models

Next, the researcher could describe a general analytic strategy that will be followed when comparing competing models. Analytic strategies have long been discussed for multiple regression analysis, and readers typically expect to read about the analytic strategy a researcher used in a complicated study employing regression analysis. In small studies, simultaneous regression—with all predictors

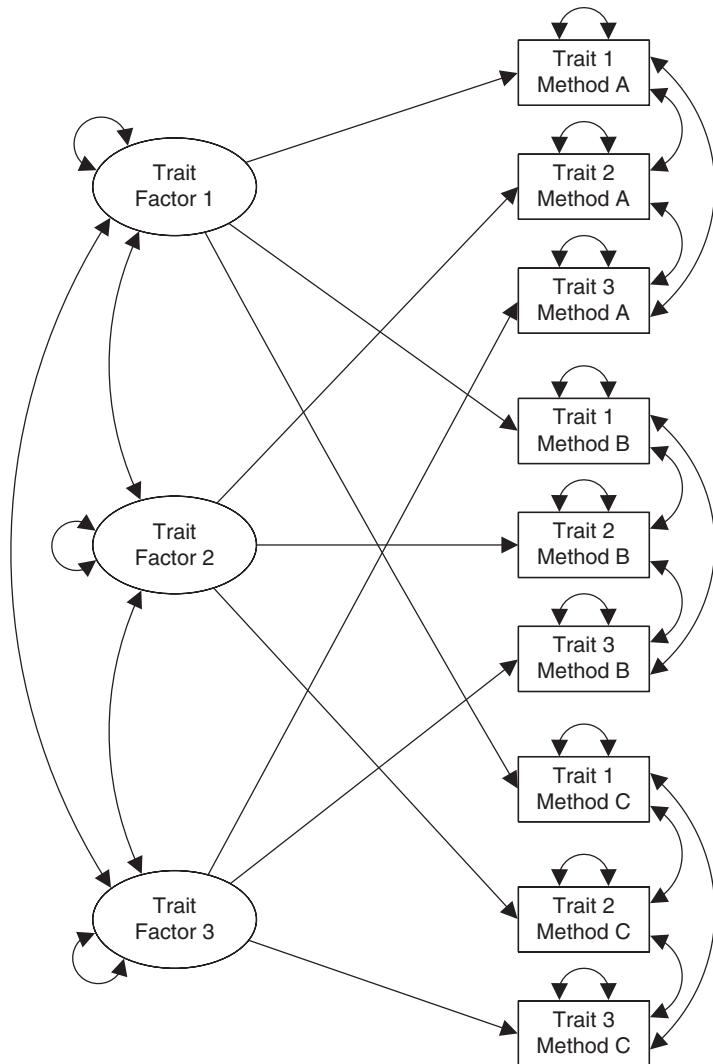


Figure 24.2 The Correlated Trait–Correlated Uniqueness (CTCU) CFA Model for MTMM Data.

included in a single regression model—is frequently used. But, if sample size is rather large and the number of potential predictors is also large, some form of theoretically or empirically driven hierarchical strategy might be used to understand the unique impacts of various sets of predictor variables.

Analogous analytic strategies should be used when evaluating the fit of a structural model to an MTMM matrix. Widaman (1985) outlined three alternative strategies, each based on a common strategy under multiple regression analysis. One strategy was akin to simultaneous regression analysis; under this approach, one might fit all of the models in the Widaman (1985) taxonomy and compare the importance of particular estimates whenever and wherever possible. A second strategy had similarity to forward selection in regression analysis. Under this strategy, one would start with the simplest model that might reasonably account for the data (i.e., a model with correlated traits and no method factors) and then add method factors and finally correlations among method factors

if these were required to explain patterns in the MTMM matrix. The third strategy was similar to methods in regression analysis in which one partials out the influence of nuisance variables before testing the primary effects of interest. Here, the researcher could begin with a model that contains only correlated method factors and then adds trait factors and finally correlations among trait factors if needed to represent the relations in the MTMM matrix. Regardless of which analytic strategy is selected, the researcher should describe and justify the analytic approach taken in the study.

6. Sampling Method and Sample

As in any empirical study, the researcher should carefully and fully describe the nature and size of the sample of participants. Studies that utilize an MTMM matrix are oriented toward evaluating measurement properties, and measurement properties (e.g., reliability) are known to vary across samples, especially if a sample exhibits restriction of range of scores on any manifest variables. Therefore, the researcher is advised to consider carefully the sample selection methods used and to obtain as representative a sample from a population as possible. Samples of convenience are to be avoided in MTMM matrix studies, because the researcher intends to provide general conclusions regarding the trait- and method-related variance in measures. Further, if any variables might be confounded in the correlations in the MTMM matrix, the effects of these should be estimated and, if large, controlled statistically. Potential confounding variables include nesting of participants within groups (e.g., students within classrooms), which can lead to failure to meet the assumption of independence of observations. With such effects, multilevel approaches to model fitting can be utilized. Or, if socioeconomic status, sex, or other background variables are related to the measures in an MTMM matrix, these variables could be included as control variables or covariates within the CFA model for the MTMM data.

The size of the sample of participants is of particular interest when structural equation modeling techniques are used. Monte Carlo studies of SEM have generally found that sample sizes of 150–200 or more should be used to estimate parameters and their standard errors (SEs) accurately if maximum likelihood (ML) estimation is used. If manifest variables fail to exhibit univariate and multivariate normality, then some more advanced methods of estimation, such as asymptotically distribution free (ADF) or weighted least squares (WLS), or robust adjustments to ML statistics (Satorra-Bentler corrections), might have to be used to ensure accurate estimation of parameter estimates and their SEs, and these methods typically require more substantial sample sizes (e.g., 1,000 or more for ADF). If a researcher has reasonable *a priori* estimates of trait and method factor loadings, along with correlations among the trait and among the method factors, the investigator should explore power analyses to ensure that s/he has sufficient power to detect the parameter estimates expected and/or to detect meaningful differences between competing MTMM models.

7. Identify and Describe Manifest Variables

Every empirical study in the social and behavioral sciences should have a clear description of the manifest variables included in the study. However, because the focus of any study featuring the use of the MTMM matrix is a full evaluation of the psychometric properties, particularly the trait and method decomposition, of manifest variables, the need to provide a careful description of all manifest variables is heightened. Most commonly, the manifest variables in an MTMM matrix study are not individual items, but are scale scores composed as the sum (or average) of multiple items. In a typical study, the researcher should separate out the description of each manifest variable in its own brief paragraph, so readers can easily identify the part of the Method section that pertains to each manifest variable.

For each manifest variable, the researcher should provide a complete description of the way in which the manifest variable score was derived. This description should include the number of items in the scale, the response scale used for each items (i.e., the number of scale points and the scale endpoints), and the way in which the score on the manifest variable was derived. The scale or manifest variable score is usually obtained as the simple, equally weighted sum of all items that comprise the scale. However, if any differential weighting or other unusual method of arriving at a total scale score for the manifest variable is used, this must be clearly described and justified. Additionally, an index of reliability, such as coefficient alpha or coefficient omega, should be reported for each manifest variable.

8. Latent Variables and Their Indicators

The manuscript should have a clear description of the manifest variables associated with each trait and method factor. As noted earlier, any MTMM matrix study should strive to have measures of at least three traits obtained using at least three methods to ensure adequate identification of all trait and method latent variables. Because the focus of an MTMM matrix study is on the measurement properties of manifest variables, the delineation of which variables load on which latent variables is usually a straightforward matter. Regardless of its often pro-forma nature, these relations should be clearly stated. Note that some researchers have formulated second-order MTMM designs (e.g., Marsh & Hocevar, 1988). To employ a second-order MTMM design, one must have multiple indicators for each of the trait-method combinations in a study. Then, first-order factors are specified as having direct effects on appropriate manifest variables, and the MTMM matrix would be found in the matrix of correlations among first-order factors. Second-order factors would then be specified as having direct effects on appropriate first-order factors. The complexity of a second-order MTMM design underscores the importance of outlining carefully the relation of manifest variables to their associated latent variables and the place of measurements within the MTMM matrix.

The researcher should also clarify and justify how the latent variables in the model were identified. Many experts on SEM recommend fixing to 1.0 the factor loading of one indicator for each latent variable. This specification is usually sufficient to identify the scale of each latent variable and allow the estimation of all additional parameters of the model. However, this choice of identification constraint usually leads to latent variables with variances that depart from unity, so covariances among trait factors and among method factors are in a metric that can be difficult to interpret. Thus, it is often advisable to fix to 1.0 the variance of all latent variables in the model, which ensures that covariances among latent variables will be scaled in the metric of correlations, making them much easier to interpret. In addition, this choice for model identification leads to the estimation of all trait factor and method factor loadings and their associated SEs, enabling a more informative set of estimates for these key model parameters.

As discussed in a later section on problems in model fitting (Desideratum 11), the CTCM model often leads to lack of convergence and/or improper parameter estimates. To counter these problems, Castro-Schilo et al. (2016) recently described systematic ways to augment the core MTMM model with additional manifest variables and thereby to improve the estimation of model parameters. The methods proposed by Castro-Schilo et al. largely mitigated analytic problems associated with the CTCM model. If researchers wish to use these new methods, the description of which manifest variables are used as indicators for which trait and method factors should be carefully outlined.

9. Analytic Methods

Many different SEM programs can be used to fit the standard structural models for MTMM data. These programs are listed in Chapter 33 in this volume. Most structural models for MTMM data are

relatively simple to specify, so virtually all SEM programs will give identical parameter estimates and SEs for these models. Still, the researcher should provide the name and version number of the program used for analyses, as SEM programs are in a continual state of revision and readers should be informed of the specific program used for analyses. The researcher should also clearly state the method of estimation used. ML estimation is the default in virtually all SEM programs and is often the method used for analyses of MTMM data. However, ML estimation rests on stringent assumptions, including the assumption of multivariate normality of manifest variables. Thus, the researcher should address whether the MTMM data meet these assumptions; for example, the researcher should report univariate and multivariate indices of kurtosis, as departures from optimal kurtosis can lead to bias in model fit statistics. If the data exhibit either platykurtosis or leptokurtosis, then some method other than ML may be more appropriate and should be explored. Furthermore, ML estimation can have difficulties in estimating small parameter estimates, and as a result certain MTMM models, particularly the CTCM model, often have poor analytic outcomes, likely due to the use of ML estimation. Recently, Helm et al. (2016) showed that Bayesian estimation of the CTCM model essentially eliminated estimation problems that occurred under ML estimation.

In addition to other technical details on analytic methods, the researcher should note clearly whether the structural models were fit to the MTMM matrix in a correlational or covariance metric. Most structural models are covariance structure models, not correlation structure models. As a result, fitting covariance structure models to correlational data can lead to inaccurate estimation of many statistics, including the overall chi-square index of fit, factor loadings, factor intercorrelations, and SEs of all parameter estimates. Even when all other statistics are estimated accurately, the SEs of parameter estimates will almost always be biased if covariance structure models are fit to a correlation matrix. Unfortunately, this matter received too little attention in prior applied analyses of MTMM data; indeed, most studies in which structural models have been fit to such data apply covariance structure models to the MTMM matrix in correlational metric. To ensure proper estimation of parameter estimates and SEs, it is strongly recommended that models be fit to MTMM data in covariance metric, and standardized estimates can still be reported to allow less complicated interpretation.

10. Missing Data and Outliers

In many types of study (e.g., longitudinal studies), missing data are more the norm than the exception. In contrast, given the fact that studies that use the MTMM matrix approach usually require only a single time of measurement, many such studies do not have missing data. Regardless, the researcher should describe the extent and the origins of any missing data in a study. Researchers often avoid missing data by using listwise deletion of participants, which leads to the dropping of participants with missing values on any manifest variable. This is not a generally recommended practice, because it leads to bias in the estimation of correlations among manifest variables. Instead, researchers should explore the use of full information maximum likelihood (FIML) estimation, in which models are fit directly to raw data matrices with missing data, or multiple imputation (MI), which analyzes and aggregates results from several datasets containing randomly varying imputed values. These approaches should lead to less bias in the estimation of the initial relations among manifest variables and then less bias in estimates of model parameters and standard errors.

The researcher should also discuss the presence of any univariate or multivariate outliers on manifest variables and how such outliers are to be handled. Outlier detection is rarely discussed in studies that use the MTMM matrix approach, yet outliers can distort relations in the MTMM matrix, so are not a trivial issue. Greater detail on outlier detection and how to handle outliers is provided in the SEM chapter (Chapter 33) so will not be discussed in further detail here. Suffice it

to say that optimal results in MTMM matrix studies will be obtained only if the influence of outliers is minimized, so researchers are strongly encouraged to use state-of-the-art methods of detecting and handling outliers.

11. Problems with Model Fitting

Studies that use SEM frequently encounter problems in the fitting of models to data. These problems in model fitting can be of several forms, including lack of convergence and the presence of improper parameter estimates. The investigator should describe how any problems in model fitting will be handled. Lack of convergence arises when the iterative estimation routines in an SEM program fail to meet the convergence criterion and therefore fail to arrive at the ML solution within a specified number of iterations. Each SEM program has a default number of estimation iterations, which are often a function of the size of the problem or of the number of estimates in the model. The output from the SEM program should indicate in some fashion why the program failed to converge, although the user might have to be especially attentive to note whether convergence was achieved or whether iteration halted because the default number of iterations was exceeded. If the default was exceeded, the user can increase the number of iterations. If this does not solve the convergence problem, the model may require re-specification because the model may not be sufficiently well identified empirically.

The presence of improper parameter estimates is the second major class of problems that often arise when fitting models to MTMM data. The most common types of improper estimates are (a) estimated correlation coefficients that fall outside the range from -1.0 to $+1.0$ and (b) negative variances. At times, the estimated correlation between a pair of trait or method factors falls outside the mathematically acceptable range from -1.0 to $+1.0$, because latent variable variances and covariances are estimated separately and no joint constraints are automatically imposed to ensure that correlations remain within mathematically acceptable bounds. If unacceptable estimates of this nature occur, the factors with unacceptably high correlations cannot be distinguished empirically, and the model must be respecified so that all model estimates are acceptable. For example, if Trait Factor 1 and Trait Factor 2 are estimated to correlate $+1.1$, one could either respecify the model so that all indicators for both Trait Factors 1 and 2 load on a single trait factor that subsumes the two domains of content or constrain the estimate of the correlation between Trait Factors 1 and 2 to fall on or within the boundary of the acceptable parameter space (i.e., fall between -1.0 and $+1.0$).

With regard to the second class of problematic estimates, the estimate of a variance parameter can be negative, and a negative variance is unacceptable because a variance is the square of the corresponding SD so must be greater than or equal to zero. Various possible bases for negative variances have been discussed (van Driel, 1978); regardless of the basis, a negative variance is unacceptable. In responding to this problem, one can either (a) fix the variance estimate to a value that is on or within the boundary (e.g., fix the variance to zero or to some small positive value) or (b) constrain the variance estimate to be greater than or equal to zero or to some small positive value. If the negative variance is a unique variance, then an appropriate model constraint may be employed. Suppose the researcher has an estimate of the reliability of the manifest variable, where r_{yy} is used to denote the reliability of variable Y. The researcher could then constrain the unique variance for the manifest variable to be greater than or equal to $s^2_y(1 - r_{yy})$, where s^2_y is the variance of variable Y. The quantity $s^2_y(1 - r_{yy})$ for a manifest variable represents its estimated error variance, which is a lower bound estimate of unique variance for the manifest variable. Regardless of how the problem is handled, a negative error variance is often a symptom of a larger problem with the model, so some form of model respecification is typically needed.

As outlined above, the Helm et al. (2016) paper on Bayesian estimation of models for MTMM data showed clearly that virtually all estimation problems encountered using ML estimation of the CTCM model can be solved by the use of Bayesian estimation. Although further work on this topic

is justified, the use of Bayesian estimation may circumvent most problems encountered in fitting models to MTMM data, so should be explored if problems arise when fitting models to MTMM data using the default ML method of estimation.

12. Descriptive or Summary Statistics

The key set of descriptive statistics for any MTMM study is the matrix of correlations among the manifest variables, along with the means and SDs of the manifest variables. Following the original presentation by Campbell and Fiske (1959), researchers usually array the measures of the t trait factors in the same order within each of the m methods. That is, the rows and columns of the correlation matrix are arranged so that the first t measures are the trait measures obtained using the first method of measurement, the next set of t measures are the trait measures obtained using the second method, and so forth. Arrayed in this fashion, the several key parts of the MTMM matrix—including the validity diagonals that contain the convergent validities—will be in the expected places that will ensure easy interpretation of the MTMM matrix by readers.

In addition to the simple presentation of a table with the MTMM matrix of correlations, the investigator should offer some summary observations on the correlations contained in the matrix. For example, Campbell and Fiske (1959) argued that convergent validities should be statistically significant and sufficiently large to encourage further research with the manifest variables. Campbell and Fiske then discussed how one should compare the validity diagonal elements to other elements in the matrix to evaluate the discriminant validity of the measures. As a result, prior to fitting structural models to the data, the investigator should describe the general levels of convergent and discriminant validity exhibited by manifest variables in the MTMM matrix. This would include noting both the general magnitude of the convergent validities and the degree to which the convergent validities tend to exceed the magnitudes of other relevant correlations and thereby exhibit discriminant validity.

13. Fit Relevant MTMM Models

The researcher should next select an analytic strategy for fitting relevant structural models to the MTMM data. Widaman (1985) discussed three general analytic strategies, the first of which is a simultaneous strategy, which involves the fitting of the entire set of models in the taxonomy of models he proposed. With four trait structures and four method structures, this would entail the fitting of 16 different structural models to the MTMM data. The investigator could supplement this set of models with additional models proposed by Marsh (1989), who incorporated a fifth method structure—the correlated uniqueness method specification—into the taxonomy proposed by Widaman (1985), or the model proposed by Eid (2000).

Rather than fitting all possible models discussed by Widaman (1985), research in a given area may be sufficiently advanced that only a subset of models should be considered. If a more restricted set of models should be of interest, one could implement one of two stepwise analytic strategies. The first of these is a forward selection strategy, in which one would first fit a model with correlated trait factors, then add to this model orthogonal method factors, and finally allow the method factors to correlate. The second approach is based on first partialing irrelevant sources of variance. Under this approach, one might first fit a model with correlated method factors (because these represent construct-irrelevant variance) and then subsequently add trait factors and correlations among the trait factors to determine whether the addition of trait factors leads to improved representation of the data over and above the estimation of method factors. The primary principle guiding the choice of an analytic strategy is this: Prior research and theory should dictate which sources of variance—trait factors and/or method

factors—might be expected to explain correlations among the manifest variables, and these considerations should, in turn, dictate the set of models to be fit to the data.

14. Evaluate Relative Overall Fit of Competing Models

Once the set of MTMM structural models are fit to the data, the researcher should evaluate the relative fit of alternative models. The Widaman (1985) taxonomy, supplemented with the Marsh (1989) method structure or the Eid (2000) model, can be employed to determine which model comparisons are legitimate when testing statistical differences in fit between models. Statistical differences in fit between two models can be tested if one model is nested within the second model. The latter, more highly parameterized model must have all parameter estimates in the first, more restricted model; the more restricted model can be obtained from the more highly parameterized model by fixing one or more parameter estimates in the latter model to zero. For example, a model that contains correlated trait and correlated method factors is a highly parameterized model, which can be designated Model 3. One could obtain one more restricted model, Model 2, by fixing correlations among method factors to be zero; a researcher could obtain a still more restricted model, Model 1, by fixing all method factor loadings to zero and thereby eliminating method factors. With this set of models, Model 1 would be nested within Model 2, and Model 2 would be nested within Model 3. Given this set of nesting relations, differences in fit between the various models can be studied.

Differences in the fit of nested models can be evaluated in several ways. The most common tool for investigating the differences in fit of nested models is to use the likelihood ratio (or chi-square difference) test. Basically, if one model is nested within another, the difference in chi-square values is distributed as a chi-square variate with degrees of freedom equal to the difference in degrees of freedom for the two models. The likelihood ratio chi-square test of a model, and any related chi-square difference tests comparing nested models, are heavily influenced by sample size. In such situations, researchers often rely on practical fit indices to evaluate the difference in fit between models. Certain practical fit indices are termed measures of parsimony-corrected absolute fit because they contain a correction or adjustment for model complexity, yet index fit for a given model without regard to other, more restricted models; among these measures, the root mean square error of approximation (RMSEA) is perhaps the most useful. Other practical fit indices are termed measures of relative or incremental fit because they involve the comparison of fit of a given model to that of a more restricted, null model. Of the relative fit indices, the comparative fit index (CFI) and the Tucker-Lewis index (TLI) appear to be optimal indices of fit. The interested reader is referred to Widaman and Thompson (2003) for a more in-depth consideration of these measures.

Finally, model comparisons can be made among models that have no nesting relations. The most common measures used for such comparisons are information indices, such as the Akaike information criterion (AIC) and the Schwarz Bayesian information criterion (BIC). For the AIC and BIC, the model with the smaller value of the index is the preferred model. These information criteria are also useful if two trait factors or two method factors are merged because the correlation between the two factors nears 1.0. Although the merging of two factors does lead to nesting of models under typical definitions of nesting, the resulting likelihood ratio difference test statistic does not follow a chi-square distribution because the correlation between factors is fixed at a boundary value (i.e., 1.0).

15. Post-Hoc Model Modifications

The investigator should clearly describe and justify any post-hoc modifications to the a priori specification of models to fit the MTMM matrix. The specification of trait factor and of method factors is a relatively simple and straightforward enterprise. However, the CTCM model is often poorly

identified empirically, so is prone to problems of lack of convergence and improper estimates. To counter identification problems, modifications to the a priori specification of a model can lead to a more well-conditioned model that converges and has acceptable estimates of all parameters. One example of a post-hoc modification is the constraining to equality of all factor loadings on each method factor (presuming all variables to be in the same metric). Many times, method factor loadings are not large, and estimation problems can arise in such situations. These estimation problems can be avoided by constraining to equality all factor loadings on a given method factor.

A second class of post-hoc modifications is the respecification of the nature of method factors. The studies by Stacy et al. (1985) and Widaman et al. (1993) exemplify principled respecifications of method factors. In the latter study, the authors used three methods—standardized instrument, day shift ratings, and evening shift ratings—of four trait dimensions of adaptive behavior—cognitive competence, social competence, social maladaption, and personal maladaption. The original method factor specification had all four measures from a given method of measurement loading on a single factor for that method, but this method failed to achieve proper estimates. Based on examination of model fit, Widaman et al. respecified the method structure with two method factors for each method of measurement—one factor with loadings from the two dimensions of competence (cognitive and social) and a second factor with loadings from the two dimensions of maladaption (social and personal)—and allowing these method factors to correlate within methods, but not between methods. This respecified model fit the data very well. Regardless of the form of post-hoc respecification, any altered specification of a model must be justified on both theoretical and empirical grounds.

16. Evaluate Fit of Optimal Model to Manifest Variables

Once a final, acceptable model is selected to represent the MTMM matrix, the researcher should carefully interpret and evaluate all parameter estimates in the model. Evaluation of estimates is usually done in several ways. First, the point estimates of the factor loadings should be noted, at least with regard to their mean (or median) value and range. Thus, the investigator might say that the trait factor loadings were moderate to strong, with a median loading of .72 and a range from .55 to .80. Similar comments can be made regarding the method factor loadings and the correlations among the trait and/or method factors.

The researcher should also evaluate each parameter estimate for statistical significance and likely population value. The critical ratio of the parameter estimate divided by its standard error yields a large-sample z statistic, and the common practice in the field is to require a z statistic to be 2.0 or greater to declare the associated parameter estimate significant at the $p < .05$ level. The investigator could also provide an interval estimate of each parameter estimate or of certain, key selected parameter estimates. An approximate 95% confidence interval can be constructed as the parameter estimate plus or minus twice its standard error, and this interval estimate is often more useful as an indication of likely population values of a given parameter than is the simple z statistic.

As discussed by Widaman (1985), certain parameter estimates in the CTCM model are directly related to the stipulations regarding convergent and discriminant validity discussed by Campbell and Fiske (1959). The following estimates are of particular interest: (a) trait factor loadings are the primary indicator of convergent validity, because they represent the relation of the trait latent variable to its indicators, so higher trait factor loadings indicate higher levels of convergent validation; (b) correlations among trait factors are the principal basis for representing discriminant validity, as the correlations among trait factors indicate how discriminable the latent factors are empirically, so correlations among trait factors that tend toward zero indicate better discriminant validity than do high correlations among trait factors; and (c) method factor loadings are the primary indicator of the how strongly methods affect or influence manifest variables, so lower method factor loadings

are preferred. The investigator should discuss these different sets of parameter estimates to provide the interested reader with a concise description of the degree to which the manifest variables in the MTMM matrix exhibit optimal patterns of convergent and discriminant validity.

17. Report Estimates of Trait and Method Variance

The investigator should compute and report the proportion of variance in each manifest variable due to trait, method, and unique factors. This is a relatively simple task if the manifest variables are in standardized, or correlational, metric. Provided that no correlations are allowed between trait and method factors and that each manifest variable loads on only its appropriate trait and method factors, then the proportions of variance explained by trait and method factors are simply the squares of the standardized trait and method factor loadings, and the sum of trait, method, and unique variance should be unity.

By default, structural equation modeling programs typically fit covariance structure models to data, and these models should be fit to covariance matrices. Unfortunately, most analyses of MTMM data involve the fitting of models to correlation matrices, as noted in Desideratum 9 above. If a researcher performs analyses in an optimal fashion and fits MTMM structural models to the MTMM matrix of covariances among manifest variables, the computation of variance due to trait, method, and unique factors is somewhat, but only slightly more complicated. That is, the variance of each manifest variable is represented in such a model as the additive sum of the square of the trait factor loading, the square of the method factor loading, and the unique variance, a total variance that usually departs from unity. Dividing each source of variance (e.g., the squared trait factor loading) by the total variance will provide the desired estimate of variance due to that source. Alternatively, the researcher can fit an MTMM CFA model to the covariance matrix and then present parameter estimates from the standardized solution. The researcher could additionally report a confidence interval for each variance estimate to supplement the point estimate of variance that is usually reported.

18. Implications of Modeling for Manifest Variables

The investigator should discuss the results of the MTMM analyses for the manifest variables included in the matrix. One of the key points made by Campbell and Fiske (1959) was that evidence for convergent validation of measures should be sufficiently strong to encourage further research in a given domain and, presumably, with the particular manifest variables in the analysis. The results of fitting structural models to MTMM data provide estimates of trait-related and method-related variance. Although no hard-and-fast rules for adequacy of trait-related variance in measures, some guidelines can be suggested. In much Monte Carlo work on factor analysis, standardized factor loadings of .4, .6, and .8 have been used to represent low, medium, and high levels of communality. After squaring these loadings, this means that researchers generally consider explained variance figures of .16, .36, and .64 to reflect low, medium, and high levels of saturation of the manifest variable with the factor. Whether these guidelines are acceptable in any particular application of structural modeling to MTMM data is a subject matter concern. Thus, in some domains, saturation of .16 with a trait factor for one or more manifest variables may be considered adequate; in other domains, minimal levels of trait factor saturation might be .25 or .30. Prior research in the area—and prior research using the manifest variables used in the current study—would be very valuable information when interpreting results.

The investigator should keep in mind the fact that, in the typical MTMM structural model, each manifest variable is influenced by two factors—one trait factor and one method factor. As a result, higher levels of factor saturation (e.g., above .60) are unlikely to be achieved routinely for

saturation with a trait factor because the trait and method factors are each attempting to explain variance of the manifest variables. Still, the guidelines listed above may prove useful when interpreting the magnitude of effects of trait factors and method factors on the manifest variables.

19. Implications of Modeling for Future Research

The researcher should also discuss the implications or impacts of the results of the current study for future research on the traits and the substantive domain. Provided that very strong convergent validation of measures was obtained (see, e.g., Stacy et al., 1985), the researcher might suggest that little is gained by inclusion of many methods of measurement in the future. Still, to avoid bias associated with any particular method of measurement, one might reasonably recommend the use of several methods of measurement in any future study in order to triangulate in the assessment of key constructs of interest.

In many areas of research in the social and behavioral sciences, researchers continue to use favorite ways of assessing constructs because these approaches have been widely used in the past. But, if MTMM studies have not been performed in a particular domain, questions may linger regarding whether the results obtained are closely related to the trait construct presumably assessed by a measure or whether method influences on the measure might represent a major contaminant on scores. Because of this, research in any domain of investigation should be supplemented with MTMM data collection to verify the amounts of trait- and method-related variance in manifest variables. Successful MTMM studies will buttress traditional measures of constructs, adding important measurement-related information to the existing literature. On the other hand, MTMM studies that yield rather low levels of trait-related variance for key manifest variables can lead to important reorientations in fields of inquiry. Regardless of the degree of success in confirming high amounts of trait-factor saturation in measures, the results of MTMM studies have important implications for research on the trait constructs and the methods of assessing these constructs, implications that should be drawn out with clarity.

20. Limitations

As a final note, the researcher should discuss any limitations of the current study. Limitations may arise at many levels. For example, use of a sample of participants from a university subject pool may result in a rather restricted sample, and the results obtained may not generalize to the population. Or, the precise manifest variables included in the study may be problematic in some ways. Because MTMM studies often include more variables than typical studies, researchers may use shortened forms of measures due to the need to assess a large number of constructs. If these shortened forms fail to exhibit high amounts of trait-related variance, the problem may be due more to the use of shortened forms with lowered levels of reliability than to the problematic nature of assessing the constructs under study.

Sample size is always a consideration when using structural equation modeling, and studies using MTMM data and models are no exception. Preferably, sample size should be rather large, generally with a minimum sample size of 150 or 200 participants, and larger samples should be sought as the size of the model increases. MTMM structural models tend to be somewhat less stable and therefore more prone to lack of convergence than do typical structural models. Therefore, a researcher should be encouraged to obtain as large a sample of participants as possible. Regardless of any limitations, the use of the MTMM approach to construct validation is still considered one of the strongest approaches one can take, and the results of MTMM studies will always represent important contributions to the research literature.

Acknowledgments

This research was supported by Grant HD076189 from the National Institute of Child Health and Human Development (David Hessl, PI), and Grant AG021029 from the National Institute on Aging (Dan Mungas, PI).

References

- Browne, M. W. (1984). The decomposition of multitrait–multimethod matrices. *British Journal of Mathematical and Statistical Psychology*, 37, 1–21.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Castro-Schilo, L., Grimm, K. J., & Widaman, K. F. (2016). Augmenting the correlated trait-correlated model for multitrait–multimethod data. *Structural Equation Modeling*, 23, 798–818.
- Eid, M. (2000). A multitrait–multimethod model with minimal assumptions. *Psychometrika*, 65, 241–261.
- Eid, M. (2006). Methodological approaches for analyzing multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 223–230). Washington, DC: American Psychological Association.
- Eid, M., & Diener, E. (Eds.) (2006). *Handbook of multimethod measurement in psychology*. Washington, DC: American Psychological Association.
- Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait–multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods*, 8, 38–60.
- Helm, J. L., Castro-Schilo, L., & Oravecz, Z. (2016). Bayesian versus maximum likelihood estimation of multitrait–multimethod confirmatory factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, 24, 17–30.
- Hubert, L. J., & Baker, F. B. (1978). Analyzing the multitrait–multimethod matrix. *Multivariate Behavioral Research*, 13, 163–179.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36, 109–133.
- Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait–multimethod matrix. *Journal of Experimental Social Psychology*, 12, 247–252.
- Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait–multimethod data. *Psychological Methods*, 7, 228–244.
- Marsh, H. W. (1989). Confirmatory factor analyses of multitrait–multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, 13, 335–361.
- Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait–multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology*, 73, 107–117.
- Millsap, R. E. (1995). The statistical analysis of method effects in multitrait–multimethod data: A review. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 93–109). Hillsdale, NJ: Erlbaum.
- Ostrom, T. M. (1969). The relationship between the affective, behavioral, and cognitive components of attitude. *Journal of Experimental Social Psychology*, 5, 12–30.
- Reichardt, C. S., & Coleman, S. C. (1995). The criteria for convergent and discriminant validity in a multitrait–multimethod matrix. *Multivariate Behavioral Research*, 30, 513–538.
- Stacy, A. W., Widaman, K. F., Hays, R., & DiMatteo, M. R. (1985). Validity of self-reports of alcohol and other drug use: A multitrait–multimethod assessment. *Journal of Personality and Social Psychology*, 49, 219–232.
- Van Driel, O. P. (1978). On various causes of improper solutions of maximum likelihood factor analysis. *Psychometrika*, 43, 225–243.
- Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait–multimethod data. *Applied Psychological Measurement*, 9, 1–26.
- Widaman, K. F., Stacy, A. W., & Borthwick-Duffy, S. A. (1993). Construct validity of dimensions of adaptive behavior: A multitrait–multimethod evaluation. *American Journal on Mental Retardation*, 98, 219–234.
- Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, 8, 16–37.

25

Multivariate Analysis of Variance

Keenan A. Pituch and Wanchen Chang

Multivariate analysis of variance (MANOVA) is a statistical model that is appropriate for both experimental and non-experimental research where associations between one or more explanatory (independent) variables and multiple outcome (dependent, response) variables are of interest. While outcome and explanatory variables may, in general, be quantitative or qualitative, this chapter focuses on the analysis and interpretation of statistical models involving only *qualitative* explanatory variables, that is, variables that are used to group the available units, typically human participants, and only quantitative outcomes. As presented here, MANOVA, used with *descriptive discriminant analysis* (DDA), is viewed as an extension of the univariate general linear model (see Chapter 1, this volume, on between-subjects ANOVA) where the purpose is to examine population differences on one or more linear composites of correlated outcome variables. The correlations among the outcomes are assumed to be due to one or more constructs that underlie the observed measures (here, “constructs” are conceptualized somewhat differently than in a structural equation modeling context; see Chapter 33 this volume). With MANOVA, composites are weighted linear combinations of the observed variable scores with the estimated weights specifically designed to maximize group separation. That is, the composite variables are created in such a way as to obtain the largest differences in group means on the composite variables. These composites are called *linear discriminant functions* and each function defines an independent construct. It is the difference between populations on these constructs that is of primary interest to the researcher.

The purpose of DDA is to identify, define, and interpret the constructs determined by the linear composites that separate the populations being compared. The careful selection of the outcome variables is thus essential for a meaningful analysis. A researcher may begin a study having some idea regarding the underlying constructs, but unanticipated constructs may be suggested by the analysis results. DDA can be used to support the researcher’s beliefs regarding the assessed constructs as well as to suggest new or unanticipated constructs that may underlie the observed outcome measures.

Additional discussions of the application and interpretation of MANOVA and DDA that are less technical can be found in Hair, Anderson, Tatham, and Black (2005), Huberty and Olejnik (2006), and Pituch and Stevens (2016). More mathematical discussions of MANOVA can be found in Anderson (2003), Johnson and Wichern (2007), and Rencher and Christensen (2012). Guidelines for preparing or evaluating studies using MANOVA with DDA are presented in Table 25.1. These guidelines are elaborated upon in subsequent sections.

Table 25.1 Desiderata for Multivariate Analysis of Variance.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. Specific research questions and the research design are explicitly stated.	I
2. An appropriate rationale for the use of multivariate analysis is provided.	M
3. Constructs of interest are discussed and the selection of the outcome variables as indicators of those constructs is presented.	M
4. In a series of preliminary analyses, data are screened and the appropriateness of applying the statistical model is verified. Any modifications of the original data, (e.g., deleting or combining measures, removing cases) are reported.	M
5. The specific version of the statistical software package used in the analyses is stated.	M, R
6. Hypothesis test results for the associations between the grouping variables and the set of constructs are presented.	R
7. The strength of the association between the grouping variables and the set of constructs is indicated.	R
8. The number of constructs responsible for meaningful group differences are determined and reported. Wilks's Λ test results, proportion of variance statistics, group mean centroids, and a plot (or plots) of the centroids should be reported for the relevant discriminant functions.	R
9. The constructs responsible for meaningful group differences are defined based on the associations between the composite and outcome variables. Standardized discriminant function coefficients should be reported.	R
10. Planned comparisons focused on group differences specified by the research questions are reported, if relevant.	R
11. Associations between grouping variables and the linear composites of the observed outcome variables defining the constructs that underlie the observed measures are presented. Unanticipated constructs suggested by analysis results are discussed.	D
12. The current findings are generalized and related to previous findings. Limitations are recognized and additional questions are raised.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

1. Research Questions and Design

The substantive questions that motivated the research study should be explicitly stated and justified based on theory, previous research findings, and/or the researcher's experiences. The research questions of interest provide the first indication as to whether a DDA is appropriate. Research questions answered using DDA examine associations between one or more grouping variables (e.g., participation in an *exercise program* emphasizing *strength, aerobic training*, or a combination of *strength and aerobic training*) and one or more composite variables that are formed from the multiple outcome variables (e.g., *heart rate, blood pressure, oxygen uptake, mood, attitude*). It is essential then, as part of the explication of the research questions, that researchers explicitly define the grouping variables and the outcome variables that are intended to define the constructs. With DDA, a construct is defined from the weights that link the outcome variables to the composite variable. The determination of those weights and their interpretation are critical steps in DDA.

The research questions under investigation then involve one or more constructs (in DDA "constructs" are often operationalized differently than in structural equation modeling; see Chapter 33, this volume). An example of a construct that cannot be easily measured by a single outcome

variable is *wellness*. This construct may be defined in many ways and might include both physical and psychological aspects. No single indicator can be expected to adequately define *wellness*, but taken together the construct can be better assessed. The appropriate weighting of these indicators is accomplished in DDA.

The number, definition, and formation of the grouping variables determine the research design and must be explicitly stated for an appropriate interpretation of results. Research questions stating or implying causal relations are, in general, best addressed through the formation of groups via the random assignment of units (e.g., college students) to the levels of the grouping variable. For example, a group of 60 college student volunteers may be recruited to participate in a study comparing the benefits of three types of exercise programs. From this group, 20 students might be assigned at random to each program. Of course, *random assignment* alone does not guarantee valid causal inference. Causal inference requires that no confounding variables are present. That is, no other variable(s) can be identified that offer a convincing alternative explanation for an observed association or difference between populations. On the other hand, when groups are formed by selecting units from existing populations, (e.g., students are identified who have self-selected an exercise program that emphasizes *strength*, *aerobic training*, or a combination of *strength* and *aerobic training*) the research design is non-experimental or correlational, and associations obtained are functional rather than causal. *Random selection* of units from the target populations, while desirable, can be difficult to achieve and not necessarily required for appropriate application of DDA. The selection process used in the study does not determine the nature of the associations found but does determine the generalizability of the relation. It is important for researchers to describe in some detail the units involved in the study in order to provide some guidance as to the appropriateness of generalizing the findings.

Research questions may reflect conditional, omnibus (main effect), or specific focused associations. Conditional associations are examined through the interaction of two or more grouping variables. Formally, only explanatory (e.g., grouping) variables may interact, and it is considered inappropriate to refer to an interaction between a grouping variable and a set of outcome variables. An omnibus relation is a main effect, examining the association between a grouping variable and the composite variables generalized across all levels of the additional grouping variables. Specific group differences are examined through focused tests or contrasts, comparing specific levels or combination of levels of a single grouping variable.

The research questions and design therefore lay the foundation for the appropriate analysis and presentation of the study's findings. The need for an explicit statement of the research questions cannot be over-emphasized.

While not the focus of this chapter, procedures other than DDA may be used along with MANOVA. Whereas the focus of DDA is on group differences involving composite variables, MANOVA may also be used when research questions involve examining group differences on each of several correlated outcomes where there is no intention of describing differences in means for composite variables, although there has been some disagreement in the field about this use. Traditionally, the primary reason for using MANOVA in this context is to provide control of the overall Type I error rate. Frane (2015) showed that MANOVA, when implemented with a modified Bonferroni procedure for follow-up tests (e.g., alpha-adjusted ANOVAs or *t* tests), can provide control of the *familywise* and *per-family Type I error rates*, while also, in some situations (e.g., limited number of outcome variables), providing greater power than using strictly univariate procedures (e.g., using only Bonferroni-adjusted ANOVAs). However, Frane's study suggested that as more outcome variables are included in the analysis, the use of strictly univariate procedures (using a Bonferroni correction) tend to provide greater power than MANOVA while also controlling for the

Type I error rates mentioned previously. So, MANOVA may be a good choice in this situation when the number of outcome variables is limited (2 or 3).

In addition, in this same context, a MANOVA-like procedure is particularly appropriate for two other situations (Heck, Thomas, & Tabata, 2013; Hox, 2010; Pituch & Stevens, 2016; Snijders & Bosker, 2012). The first is when a researcher wishes to test whether group mean differences are the same or differ across similarly scaled dependent variables. In an experimental setting, for example, an investigator may learn if treatment effects are stronger for some outcomes than others, which may suggest revising the nature and/or implementation of the intervention. The second situation where a multivariate procedure may be needed when the interest focuses on individual outcome variables is when response data are partially missing, that is, when participants provide response data for at least one outcome variable but have missing data on other responses. In this case, a multivariate analysis procedure can be used that employs *maximum likelihood* estimation, which is one of two methods that are considered to be superior missing data treatments (with the other being *multiple imputation*). Compared to a strictly univariate approach, the multivariate procedure can provide for better estimation of effects and increased power in the presence of incomplete response data. Note though that for these two cases (i.e., testing for equivalence of group differences across multiple outcomes and incomplete response data), MANOVA, as traditionally implemented, could not be used easily, or, at all. Instead, the multivariate analysis suitable for these situations is known as multivariate multilevel regression or linear mixed modeling. While these are useful applications that require a multivariate analysis, this chapter will focus on MANOVA with DDA used as the follow-up analysis. Readers interested in this alternative multivariate approach may consult the references listed in this paragraph.

2. Rationale for a Multivariate Analysis with Descriptive Discriminant Analysis

Researchers using DDA should provide suitable rationale for using this procedure. Note though that the inclusion of multiple correlated outcome variables is a necessary but insufficient justification for choosing DDA. Rather, an appropriate justification for DDA is an interest in examining group differences on one or more constructs defined by linear composites of the observed outcome measures. Attaining a parsimonious description of group differences, via the formation of the composite variables, is an inherent part of the rationale.

3. Constructs and the Selection of Multiple Outcomes

Because the purpose of conducting a DDA is to identify the construct(s) that separate two or more populations, the selection of outcome variables to be included in the analysis is a critical step in the planning of the inquiry. A rationale for including the selected variables should be provided. We discourage researchers from including outcome variables simply because they are easily obtainable or happen to be available at the time of data collection. Inclusion of such variables can reduce statistical power, add little to the understanding of population differences, and make the interpretation of the results much more difficult. Outcome variables selected for analysis should cluster conceptually into one or more groupings. These groupings should reflect the constructs that the researcher believes to be relevant for the determination of group separation with a single grouping variable or in the case of multiple grouping variables, population differences. The final analysis of the selected variables may or may not support the researcher's belief. Discovering unanticipated constructs as suggested by analysis results may be among the most interesting findings of the study.

4. Preliminary Analyses

While a researcher may be tempted to proceed directly to answering the stated research questions by testing specific hypotheses and identifying constructs, such temptation should be avoided. Before such results can be examined in depth, it is important for researchers to examine basic characteristics of their data and report their findings. Before comparing groups, the data within groups must be examined first. Data should be examined to determine if they are complete (i.e., there are no missing observations or scores) and whether they include unusual or outlying observations. Within each group, score reliability for each measure (see Chapter 29, this volume) and the shape of the data distributions should also be examined. The results of these analyses should be reported and any action taken to correct or transform the original data must be reported, along with a discussion of how missing data were treated. The results of these analyses might justify deleting some measures, or perhaps when sample sizes are relatively small (e.g., the number of outcome measures is similar to or greater than the error degrees of freedom) or if after examining the within-group correlation matrices it might be judged that too many highly related outcomes are included in the data set, it may be desirable to combine measures through a principal components analysis (see Chapter 8, this volume).

Once the researcher is satisfied that the data adequately represent the variables of interest within the groups, some preliminary comparisons among groups is appropriate. For each group the sample size, outcome means, and standard deviations must be reported. Initial insight into the constructs underlying the outcome variables and support for justifying the use of a multivariate analysis can be gained by reporting and interpreting the pooled within-group correlation matrix. Note that this matrix is *not* the total sample correlation matrix among outcome variables ignoring group membership, as ignoring group membership can result in spuriously high or low correlations among variables because of differences among group means. If the outcome variables are uncorrelated, then multiple univariate analyses would likely be more appropriate than DDA.

Univariate analysis of variance comparing outcome means may also be useful in providing an initial understanding of group differences. As a preliminary analysis, the univariate *F*-tests can provide an indication of the relation between individual outcome variables and the grouping variable(s). This analysis is for description only and provides insights similar to those gained by examining the pooled-within-group correlation matrix. Further, if there is no indication that an individual outcome variable is related to the grouping variable(s) the researcher might consider dropping that measure from further analyses. Because this is a preliminary analysis, an increased risk of a Type I error might be tolerable and a Bonferroni-type adjustment would not be necessary.

The within-group covariance matrices should be examined and compared across groups. Such a comparison is particularly important when group sample sizes are substantially different from each other (say, by a factor of 2). A statistical test for covariance equality (e.g., Box, 1949) may be used to formally compare the matrices, but these tests tend to be overly sensitive to small departures from covariance homogeneity because (1) they are sensitive to distributional non-normality, (2) they involve a large number of degrees of freedom, and (3) the number of variances and covariances being compared is generally large. An alternative strategy to a formal statistical test is to compare the log-determinants for the separate group-covariance matrices. The determinant of a covariance matrix is referred to as the *generalized variance*, a measure of total variance in the set of outcome variables, and taking the logarithm helps to put that value on a useful metric for comparison. Comparing the individual group log-determinants along with the log-determinant of the pooled covariance matrix provides an indication as to whether the assumption of equal covariance matrices is reasonable. If the determinants are in the same “ballpark” the researcher

may be justified in pursuing the multivariate analysis. The correlation between group size and the log-determinant of the group-covariance matrix may also be examined. A strong negative correlation (e.g., larger groups have smaller log determinants) would indicate a liberal MANOVA test for the association between the grouping variable and the linear composites while a positive association would indicate a conservative hypothesis test. While a violation of the covariance homogeneity assumption is less serious for MANOVA hypothesis tests when sample sizes are approximately equal, covariance inequality is still a potentially serious problem when attempting to identify the constructs underlying the observed outcome measures even when sample sizes are equal. If the researcher judges that covariance matrices are problematically unequal, several options exist. A statistical test for comparing the populations on the outcome measures that does not assume equal covariance matrices is available (e.g., Johansen, 1980; Yao, 1965). Alternatively, the researcher might examine the covariance matrices within groups and identify one or two variables that contribute disproportionately to the covariance inequality. It might be reasonable to analyze these variables separately and continue the DDA with a slightly smaller set of outcome variables, assuming that the analysis of the remaining variables is still meaningful. Still another alternative might be to reduce the number of groups being compared. Perhaps there are several subsets of the grouping variable that do have similar covariance matrices but the subsets' covariance matrices deviate greatly from each other. Separate multivariate tests might make sense to compare those groups that have similar covariance matrices. If only one group differs from the others, focused tests involving that group using the Yao (1965) procedure may be used to address the research questions.

5. Computer Software

Both SPSS and SAS computing software packages include several analytic procedures that are useful for carrying out a MANOVA analysis with DDA. Note that a single procedure may not provide all of the results needed for a complete DDA. Further, these procedures are revised on occasion. As such, the version of the software program used by the data analyst should always be reported.

6. Examining Overall Group Differences for Single and Multi-factor Designs

Once the researcher is satisfied that the necessary data conditions for MANOVA have been reasonably well satisfied, the analyses to answer the researcher's stated questions may proceed. Answers to the research questions may be stated as hypotheses, and these hypotheses are tested statistically using the MANOVA model. For example, a hypothesis in the null form for a single-factor MANOVA might be stated as: There is no association between participation in an *exercise program* (*strength training*, *aerobic training*, combined *strength and aerobic training*) and *wellness* related constructs as measured by five outcome measures (*heart rate*, *blood pressure*, *oxygen uptake*, *mood*, *attitude*). While a strictly univariate analysis would compare the populations represented by the available groups on each individual outcome measure separately, MANOVA tests the hypotheses about constructs (e.g., *wellness*) by simultaneously comparing group means on the composite variables formed as part of the analysis procedure. A given group mean for a composite variable is called a *mean centroid*, and the overall MANOVA null hypothesis could be stated as: There are no differences in the mean centroids for any of the composite variables for the *strength*, *aerobic training*, or combined *strength and aerobic training* groups.

Conditional tests. If multiple grouping variables are included in a factorial research design, it is likely that the research questions include inquiries into group differences among levels of one

grouping variable conditioned or depending on the level of a second grouping variable. These questions should be addressed first. For example, consider a 3×3 factorial design to investigate the association between three *exercise programs* (e.g., *strength*, *aerobic training*, combined *strength and aerobic training*), three levels of *exercise intensity* (e.g., *twice a week*, *three times a week*, or *daily*), and *wellness* as measured by several physiological and psychological measures (e.g., *heart rate*, *blood pressure*, *oxygen uptake*, *mood*, *attitude*). One research question may ask whether the association between *exercise program* and the *wellness* constructs varies for different levels of *exercise intensity*. This question is one of an *interaction* between the two grouping variables. Alternatively, several research questions might ask about the association between one of the grouping variables and the constructs at each (or at a specific) level of a second grouping variable. For example, a researcher may be interested specifically in the association between *exercise program* and *wellness* constructs for each level of *exercise intensity*. These alternative questions are referred to as *simple effects*. Both interaction and simple effect questions are conditional, meaning that an association between a grouping variable and the constructs is presumed to vary across specific levels of a second grouping variable.

Generally, when multiple grouping variables are included in a factorial design the interaction tests precede tests of simple effects but that need not be the case. The number of interactions that may be tested is determined by the number of grouping variables being considered and the meaningfulness of such analyses. However, all interactions need not be tested. For example, interactions involving more than three grouping variables are often difficult to interpret and often lack statistical power. In those cases, the sum-of-squares associated with higher order interactions may be pooled with the within-group or error sum-of-squares.

In addition, if an interaction is statistically significant, the factorial design may be simplified into a series of simpler designs. For example, given a statistically significant interaction between *exercise program* and *exercise intensity*, the research questions may focus on the relationship between *exercise program* and the *wellness* constructs for each level of *intensity*. Or, the comparison of two specific *exercise programs* at each level of *exercise intensity* might be examined. These are questions that are answered with tests of simple effects. Three points can be made regarding simple effects. First, simple effects are simplifications of a factorial design. They are an alternative conceptualization of the linear model that combines the effect of a grouping variable with an interaction effect. A three factor design can be simplified to three two-factor designs; a 3×3 factorial design can be simplified to two sets of three one-factor designs. An interaction test need not precede the simplification of a factorial design. The research questions raised may call for the simplification of the design.

Second, a simple effect resulting from the simplification of a two factor design to several one factor designs is not equivalent to a MANOVA with one grouping variable. That is, a simple-effect hypothesis test may use the pooled error sum-of-squares across all groups in the factorial design when computing the test statistic of interest and determining the constructs underlying the outcome variables, whereas a MANOVA for a single grouping variable pools the error sum-of-squares for only those groups being compared. If the MANOVA assumptions are met, particularly, homogeneity of covariance matrices, a simple effect test that uses the error term from all groups in the factorial design will provide more statistical power, as such a test will have greater error degrees of freedom and therefore greater sensitivity to identify population differences.

Third, when simple effects are considered, the researcher may wish to adjust the criterion used to judge statistical significance to limit the overall Type I error rate. While a good argument may be made for not making any adjustments, the problem of increasing the risk of a Type I error across all of the tests conducted should at least be addressed.

Omnibus or main-effect tests. An omnibus or main-effect test examines the association between the composite variables and a grouping variable generalized across all levels of

additional grouping variables. If an interaction is present, such tests, although valid, are not likely to be meaningful. However, if *exercise program* and *intensity* do not interact, a more general question on the association between *program* and *wellness* can be addressed. That is, across all levels of *exercise intensity* is there an association between *exercise program* and the *wellness* constructs? The answer to this question requires the simultaneous comparison of all three *exercise programs*. If overall group differences are found to be statistically significant, the results do not identify which program(s) best promotes *wellness*. Specific comparisons are needed and are discussed in Desideratum 10.

In a factorial study, when the number of observations per group is unequal and disproportional, the research design is *nonorthogonal*. In this case, the hypothesis tests on the grouping variables in the model are not independent of one another. For the results to be interpreted appropriately it is essential that the method used to compute group means be explicitly stated. For example, in a two-factor design a marginal column mean may be computed as the average of all observations in a column or as the average of the cell means in the column. The first approach is a *weighted* (also referred to as *type I* or *hierarchical* sum-of-squares) approach, where each cell mean in the column is weighted by the ratio of the number of units in the cell to the number of units in the column. Thus, the contribution a cell mean makes to the calculation of the column mean is in proportion to the number of observations in the cell. The second approach is an *unweighted* (also referred to as *type III* or *regression* sum-of-squares) approach where cell means within a column contribute equally to the calculation of the column mean, regardless of group size. The weighted and unweighted approaches for computing marginal group means test somewhat different hypotheses but both are valid. Unless the method used to compute the marginal means is explicitly stated, it is not clear what associations are tested (Carlson & Timm, 1974; Pendleton, Von Tress, & Bremer, 1986). Generally, in a multi-factor nonorthogonal research design the unweighted approach is preferred when the inequality in sample sizes is unintended and do not represent true differences in population sizes but rather reflect a random loss of participants. This may occur in experiments when an equal number of participants are randomly assigned to each condition but participants are lost due to reasons unrelated to the conditions studied (e.g., illness). The unweighted solution is also preferred because the effect of each explanatory variable can be tested after controlling or considering the effect of the additional factors and interactions in the model. Occasionally, the weighted solution might be used if the differences in group sample sizes are proportional to differences in population sizes they represent.

Four different multivariate test criteria have been developed and are provided by computer software to evaluate interactions, simple effects, main effects, and focused hypothesis tests. The SPSS MANOVA procedure reports these criteria as Wilks, Pillai's, Hotellings, and Roy's while the SAS and SPSS GLM procedures report these criteria as Wilks's lambda (Λ), Pillai's trace, Hotelling-Lawley trace, and Roy's greatest root. These criteria provide exact tests of hypotheses with identical *F* and significance values when the hypothesis degrees of freedom equal one. When the hypothesis degrees of freedom exceed one, the four criteria yield slightly different results but generally reach the same conclusion. While there is no consensus as to which criterion is best, Rencher and Christensen (2012) recommended Roy's criterion but only if the outcome variables represent a single construct when MANOVA assumptions hold. While they also recommend Pillai's criterion when multiple constructs are presumed to be present, they note that Wilks's criterion, due to historical usage and its ability to assess dimensionality (as discussed in Desideratum 8), will likely remain as the dominant multivariate test criterion. Researchers should be explicit regarding the MANOVA criterion used by stating its numerical value, $F(df_{\text{num}}, df_{\text{den}})$ value, and *p* value (e.g., Wilks's = .765, $F(2,23) = 3.529$, *p* = .046).

7. Strength of Association for Overall Group Differences

Using a multivariate test mentioned in the previous section, researchers might infer whether an observed association between a grouping variable and the set of composite variables appears to reflect a true association or whether the association can be explained by sampling error. Given that trivial group differences can be declared as statistically significant, researchers are often urged to report measures of effect that describe the strength of association between the grouping variable and composite variables. A number of overall multivariate effect size measures have been suggested, and several are discussed in Grissom and Kim (2012). For contexts when the grouping variable has only two levels and situations involving more than two levels but where two or more of the levels are compared using a pairwise or complex contrast, the Mahalanobis D statistic can be useful. Mahalanobis D is a measure of distance between two mean centroids that is analogous to Cohen's standardized mean difference statistic, d .

When three or more levels of a grouping variable are examined simultaneously, a measure of association between the grouping variable and the set of composite variables may be useful. Several multivariate measures of association have been proposed with the most popular ones being associated with each of the multivariate test criteria discussed in the previous section (see Huberty & Olejnik, 2006, pp. 62–65; Grissom & Kim, 2012). While the SAS computing package does not report these multivariate measures of association, the SPSS MANOVA procedure computes them and refers to them using the same label as the test criterion (i.e., Pillai's, Hotelling's, Wilks'). The GLM procedure in SPSS refers to these measures of association as partial eta squared. While the four measures of association are identical for multivariate comparisons of two groups or for complex one- df contrasts involving multiple groups, they provide slightly different estimates of association when multiple levels of a grouping variable are assessed simultaneously. Kim and Olejnik (2005) and Grisson and Kim (2012) discussed the different definitions of effect associated with these measures of association. Also, these measures tend to overestimate the strength of association. An adjustment for the overestimation provided by the effect estimated by using Pillai's test criterion was suggested by Serlin (1982) based on an adjustment for the squared multiple correlation in a multiple regression context by Ezekiel (1930). Kim and Olejnik (2005) showed that Serlin's adjustment also works well for measures of association based on Wilks's and Hotelling's test criteria, and Grissom and Kim (2012) suggest that adjusted effect size measures be used. While it may seem natural that the multivariate measure of association reported should be consistent with the researcher's multivariate test criterion choice, Grissom and Kim suggest this linkage is unnecessary. For a measure of overall strength of association between an effect and a set of composite variables, Grissom and Kim (2012) suggest use of the effect size associated with the use of Wilks's Λ along with Serlin's adjustment, although other effect size measures have merit. Researchers should report which effect size measure(s) they use.

A practical difficulty with multivariate measures of association lies in the interpretation of the computed values. In the case of univariate analyses, guidelines have been suggested by Cohen for interpreting d , eta-squared, omega-squared, and R^2 . Further, these statistics have been increasingly reported in the empirical literature and have become familiar to applied researchers. In the case of MANOVA, while rough characterizations of effect have been proposed for some measures (Cohen, 1988; Stevens, 1980; Steyn & Ellis, 2009), multivariate effect size measures do not appear to be consistently reported. Another problem with multivariate effect size measures is that they are influenced by the number and choice of outcome variables included in the study. As such, it is not clear how the strength of association should be characterized for MANOVA.

8. Determining the Number of Constructs Responsible for Group Differences

A statistically significant association between a grouping variable and the outcome variables suggests that there are group differences for at least one construct. However, it is not clear on the basis of an omnibus test, such as those described in Desideratum 6, how many constructs are responsible for group differences or what variables define these construct(s). Here, the determination of the number of constructs is considered while the next desideratum considers the process for defining them.

The number of constructs or composite variables that can be estimated is determined by the degrees of freedom ($J - 1$) for the grouping variable with J levels or the number of outcome variables (p), whichever is smaller. Generally, there are fewer between-group degrees of freedom than outcome variables. For example, when there are only two levels of a grouping variable or when planned comparisons or focused tests of pairwise or complex contrasts are of interest, the between-group degrees of freedom equals one. Regardless of the number of outcome variables included in the analysis, only one composite variable is formed. For a two factor ($J \times K$) interaction, the number of degrees of freedom is the product of the degrees of freedom for each factor [$df_{JK} = (J - 1)(K - 1)$], yielding as many composites (if sufficient outcome variables exist). Each composite variable represents a separate, independent construct that underlies the observed outcome variables. Typically, not all of the estimated constructs are meaningful or associated with group differences.

The number of constructs responsible for group differences can be determined using a two-step process. In the first step, a series of statistical tests are conducted using Wilks's Λ . If the overall association between a grouping variable and the set of composite variables is statistically significant with this criterion, this result suggests that at least one construct is responsible for group differences. To determine whether additional constructs separate groups, the variation associated with the first construct is removed and a modified Wilks's Λ is then used to test whether group differences are present on any remaining constructs. If this second test is statistically significant, it suggests that group separation exists for at least two constructs (given that the initial test was significant). This process continues by further partitioning or adjusting Wilks's Λ and stops when Wilks's Λ is no longer statistically significant.

The second step in determining the number of constructs that are responsible for meaningful group differences is to consider effect size measures associated with each of the multiple composite variables. One effect size measure is the square of the canonical correlation. Like an eta-squared measure in ANOVA, this measure represents the proportion of variation in a given construct that is between groups. A second effect size measure captures the proportion of the total between-group variation that is due to each construct. Given that the first construct, due to the way the composite variables are formed, is guaranteed to have the greatest proportion of between-group variation, it is especially important to examine the proportions that are due to each of the other composite variables. Of course, not all statistically significant composite variables are strongly associated with between-group differences. Guidelines for defining meaningful proportions of variation for these measures do not exist but larger values associated with these measures reflect stronger group differences. In addition, the original research questions and the earlier justification given for variable selection should be of help in the decision process.

In addition, examining the values of the group centroids and a plot of these values (when multiple composites are formed) should also be used to identify the number of composite variables responsible for meaningful group differences. A given group centroid is a group's mean score on a composite variable. Often, these centroids are scaled by software so that the grand mean across all groups is zero and the standard deviation is 1 for each composite variable. So, if the centroid for group 1 is .70 and for group 2 is -.30, the difference in group means for this composite variable

is $.7 - -.3 = 1$ standard deviation, which, in general, is often considered a large group difference. When multiple composite variables are present, software can readily generate plots of the group centroids, with each plot including the centroids for 2 composite variables for each of the groups. These plots allow a researcher to assess visually if large group differences appear to be present. Thus, for all statistically significant composite variables, the numerical values of group centroids should be reported along with, when applicable, a plot (or plots) of the group centroids.

Note, also, that the composite variables obtained in this procedure are weighted linear combinations of the outcome variables that represent underlying constructs and are useful to *describe* how groups differ. They should *not* be used for predictive or classification purposes. Because some computer software programs do not clearly distinguish between descriptive and predictive discriminant functions, researchers may confuse them and sometimes refer to them interchangeably. Briefly, predictive discriminant (or classification) functions (PDF) differ from composite variables formed in DDA in (1) purpose, (2) number, (3) calculation, (4) equal covariance matrices requirement (DDA requires it, PDF do not), and (5) application. Researchers should be clear that when presenting MANOVA results, the composite variables are interpreted as descriptive and not predictive.

9. Defining Constructs

Once the number of composite variables responsible for group differences has been identified, the constructs measured by the linear composite variables need to be defined. Constructs are believed to be real, yet are unobservable. The outcome variables, on the other hand, are observable and interpretable. In DDA, the definition of a latent construct relies on the associations between a given composite variable and the outcome variables (here, “latent constructs” are conceptualized differently than in structural equation modeling; see Chapter 33, this volume).

One method that is sometimes used involves examining the bivariate correlations between each outcome and each of the significant composite variables, with the correlation being called a *structure r*. While the sign for *r* is arbitrary, it must reflect the theoretical direction of the association between the construct and each outcome variable. For example, in measuring *wellness*, three variables—*heart rate*, *blood pressure*, and *oxygen uptake*—might be expected to define the physical aspect of *wellness*. The signs of the structure *r* values should reflect a theoretically reasonable pattern. For example, one might expect *heart rate* and *blood pressure* to be related to a common construct in the same direction while *oxygen uptake* should be related in an opposite direction (e.g., low heart rate and low blood pressure are associated with *wellness* while high oxygen uptake is associated with *wellness*). In this case, if all structure *r* values were positive or all negative, this pattern might be difficult to explain and could raise questions regarding the nature of the construct being measured. A relatively high value for $|r|$ suggests that the latent construct shares common variance with the outcome variable. When several outcome variables have relatively high $|r|$ values, the latent variable shares variance with these outcome variables and may be defined by what the outcome variables have in common. Using the structure *r* values, then, the construct is defined as those characteristics that are common to the outcome variables that are correlated with the composite score, although we note important limitations associated with using the structure *r* values below.

A second method that can be used to define a construct is to use standardized discriminant function weights, which like multiple regression coefficients, describe the unique association between a given outcome variable and the construct being defined. While raw score weights are available, their numerical value is determined to a great degree by the scale or variance of the outcome variable: generally, the greater the variance the smaller the raw discriminant function weight. Standardizing each of the outcome variables creates a common scale for all outcome variables and consequently

makes it possible to compare the weights associated with the outcome variables. A construct is generally interpreted based on the variables that have the greatest standardized weights, with the sign of the coefficient, as with the structure r , also being important when assigning meaning to the construct. Analogous to the multicollinearity issue in multiple regression, the values of the standardized weights are influenced to some extent by the degree of correlation among the outcome variables. If some outcome variables are highly related, the unique contribution those variables make will be small and consequently the weights will be small. Variables that have a bivariate correlation with a composite may not be associated with the composite when the intercorrelations among the outcomes are taken into account.

While there are different preferences on whether the bivariate correlations or the standardized coefficients should be used to define constructs, there are two concerns with using the structure r values. First, the structure r values are highly related to the univariate F statistics for testing group differences on each of the outcome variables, where the correlations among outcomes is not considered. That is, the absolute value of the structure r will be large for an outcome variable that has a large computed univariate F statistic. This relation between r and F has been interpreted to mean that the structure r is not a multivariate statistic because it only shows how a given variable by itself separates groups, much like a univariate analysis. As a result, some believe that structure coefficients should not be used to define the constructs (see Rencher & Christensen, 2012, pp. 300–301). Second, simulation research conducted by Finch and Laking (2008) and Finch (2010) determined that more accurate identification of the outcome variables that are related to a composite variable is obtained by use of standardized rather than structure coefficients. In particular, Finch found that use of the structure coefficients too often resulted in finding that an outcome variable is related to the composite variable when it is in fact not related (analogous to a Type I error). Finch concluded that using structure coefficients to identify important discriminating variables “seems to be overly simplistic, frequently leading to incorrect decisions regarding the nature of the group differences” (p. 48). Finch and Laking also noted that one weakness associated with the use of standardized coefficients occurs when a composite variable is related to only one outcome variable (as opposed to multiple variables). In this “univariate” type of situation, they found that use of standardized coefficients too often suggests that the composite is (erroneously) related to another discriminating variable.

The standardized discriminant function weights should be reported when interpreting the results of MANOVA, and the structure r values may be reported. These two types of coefficients address different issues regarding the identified constructs underlying the variable space created by the outcome variables. A structure r provides an index for relating an outcome to a composite score, and a standardized weight reflects the unique contribution an individual variable makes to the composite score. Given the problems associated with use of the structure coefficients, researchers are advised to use the standardized coefficients to give meaning to the constructs responsible for group separation, especially if each construct is presumed to be defined by two or more outcome variables (which should generally be the case in a well-planned study).

A related issue is the idea of identifying the outcome variables responsible for group separation in a general sense, or variable importance. Huberty and Olejnik (2006) suggested ranking outcome variables based on the numerical value of Wilks's Λ for group differences when the i th variable is deleted. Small values for Wilks's Λ indicate a strong association between the grouping variable and the construct. Thus, a relatively large Wilks's Λ value upon deletion of a variable indicates a relatively important variable. As such, variables with the largest deleted Wilks's Λ values would be judged as the most important variables. Rencher and Christensen (2012) noted, though, that unlike the standardized coefficients these test values cannot be used to identify the outcome variables that are associated with a specific composite variable. Rather, this procedure can be used to identify the

outcome variables that are important to group separation generally for all of the composite variables combined. However, they note that if group separation is largely due to the first composite variable, this procedure will generally rank outcome variables in the same order as obtained by use of the standardized coefficients.

10. Planned Comparisons

Examining conditional and omnibus associations is often done in MANOVA analyses. However, as in univariate designs (see Chapter 1), a researcher may have a limited set of theory-derived contrasts that are of interest, and such planned comparisons may be conducted when multiple outcomes are present. These focused tests involve specific pairwise and/or complex contrasts between specific levels of a grouping variable. Thus, instead of comparing all groups simultaneously, a subset of comparisons might be justified. For example, a pairwise comparison might be conducted to learn if the *strength* and *aerobic training* groups differed on the set of *wellness* variables. In addition, if one wished to learn whether the combined *strength and aerobic training* group had greater mean *wellness* than the other groups, a complex contrast could be conducted that compares the combined *strength and treatment* group to the average of the other two groups on the set of outcome variables. As was stated earlier but worth repeating, in many research contexts after determining that the data meet the necessary model conditions, it may very well be the case that the only analyses needed are those accomplished through the analyses of planned comparisons.

Note that the results of a specific contrast can be substantially different than the results from omnibus analyses. First, while the omnibus analysis may result in the identification of more than one construct, a contrast has only one degree of freedom and consequently only one construct may be identified. Second, with omnibus tests, the separation among all levels of the grouping variable is considered simultaneously. Thus, all levels of the grouping variable are compared on constructs defined by the same outcome variables. With contrasts, on the other hand, the outcome variables that define the construct which separates the groups being compared can differ depending on which groups are being compared. For example, the *strength* and *aerobic training* groups may differ on a construct identified as *physical wellness* but the *strength* and the combined *strength and aerobic training* groups might differ on a construct identified as *psychological wellness*. And third, if variable importance is of interest, the “most important” variables that separate all levels of the grouping variable can be different from the “most important” variables that separate the groups identified in the contrast.

Simultaneous comparisons of multiple levels of the grouping variable and specific comparisons through contrasts provide useful information regarding differences among the populations studied. While a priori planned comparisons may likely be conducted without need for any omnibus tests, it is possible that both omnibus tests and planned comparisons may be useful in the same study (Huberty & Olejnik, 2006, pp. 68–69).

11. Answering Research Questions

Following the analyses, the results should be summarized with respect to the research questions stated at the beginning of the article. The original research questions were introduced in the context of theory, practice, and/or previous research so the current findings must again be presented within those contexts. Outcome variables were chosen to reflect anticipated constructs and the extent to which the results support those constructs should be discussed. Equally important, the emergence of unanticipated constructs, if suggested by analysis results and meaningful given the phenomena being investigated, must be discussed. If relevant, differences in construct formation obtained in omnibus and focused tests between specific populations might also be highlighted.

12. Generalizing Findings

It is important that the researcher relate the current results to those previously reviewed. Both consistencies and inconsistencies with theory and previous research should be highlighted. Inconsistencies merit additional discussion with possible explanations. All research studies impose a number of limitations on the scope and execution of the study. These limitations should be made explicit. While care must be given not to overstate the implications of the current findings, it is also important not to minimize the contribution of the current findings as well. Being overly cautious with the interpretation can also be a serious mistake. Finally, directions for future research should be explicitly stated, as good research often introduces more questions than it answers.

References

- Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). New York: Wiley.
- Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, 36, 317–346.
- Carlson, J., & Timm, N. (1974). Analysis of nonorthogonal fixed effects design. *Psychological Bulletin*, 81, 563–570.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Ezekiel, M. (1930). *Methods of correlational analysis*. New York: Wiley.
- Finch, H. (2010). Identification of variables associated with group separation in descriptive discriminant analysis: Comparison of methods for interpreting structure coefficients. *The Journal of Experimental Education*, 78, 26–52.
- Finch, H., & Laking, T. (2008). Evaluation of the use of standardized weights for interpreting results from a descriptive discriminant analysis. *Multiple Linear Regression Viewpoints*, 34, 19–34.
- Frane, A. V. (2015). Power and type I error control for univariate comparisons in multivariate two-group designs, *Multivariate Behavioral Research*, 50, 233–247.
- Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York: Routledge.
- Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (2005). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
- Heck, R. H., Thomas, S. L., & Tabata, L. N. (2013). *Multilevel and longitudinal modeling with IBM SPSS* (2nd ed.). New York: Routledge.
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.
- Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis* (2nd ed.). New York: Wiley.
- Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika*, 67, 85–92.
- Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
- Kim, S., & Olejnik, S. (2005). Bias and precision of measures of association for a fixed-effect multivariate analysis of variance model. *Multivariate Behavioral Research*, 40, 401–421.
- Pendleton, O. J., Von Tress, M., & Bremer, R. (1986). Interpretation of the four types of analysis of variance tables in SAS. *Communications in Statistics: Theory and Methods*, 15, 2785–2808.
- Pituch, K. A., & Stevens, J. P. (2016). *Applied multivariate statistics for the social sciences* (6th ed.). Mahwah, NJ: Erlbaum.
- Rencher, A. C., & Christensen, W. F. (2012). *Methods of multivariate analysis* (3rd ed.). New York: Wiley.
- Serlin, R. C. (1982). A multivariate measure of association based on the Pillai-Bartlett procedure. *Psychological Bulletin*, 91, 413–417.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles, CA: Sage.
- Stevens, J. P. (1980). Power of the multivariate analysis of variance tests. *Psychological Bulletin*, 88, 728–737.
- Steyn, H. S., Jr., & Ellis, S. M. (2009). Estimating an effect size in one-way multivariate analysis of variance (MANOVA). *Multivariate Behavioral Research*, 47, 106–129.
- Yao, Y. (1965). An approximate degrees of freedom solution to the multivariate Behrens–Fisher problem. *Biometrika*, 52, 139–147.

26

Nonparametric Statistics

Michael A. Seaman

The term “nonparametric statistics” refers to a large set of analytic methods that do not rely on distributions with specified parameters. When using traditional parametric methods, the parameter of interest (e.g., the mean) is also a parameter of the underlying population distribution (e.g., a normal distribution), but with nonparametric statistics there is no assumption that the population distribution has a form that depends on the parameter of interest. Other terms used for these methods are “distribution-free” or “assumption-reduced” methods. In brief, these are methods that researchers can use when they are uncertain about the characteristics of the population(s) from which their data are sampled.

Most of the chapters in this book focus on a specific analytic technique. By contrast, this chapter is about a large set of techniques. Indeed, for many of the methods discussed in this book a researcher could instead choose an analogous nonparametric method. It can be a daunting task for a reviewer to become acquainted with such a large sub-discipline. Fortunately, the nonparametric methods used most in the social sciences cluster into two main groupings, and there are common characteristics within each grouping. A reviewer who becomes familiar with these characteristics can address most of the potential problems that might be present in a nonparametric analysis.

The first group of methods is based on using the ranks of the observations. Further, it relies on permutations of the sampled data and associated hypotheses tests called *permutation tests*. Researchers use such tests for the same purposes as those who use common parametric techniques, such as the *t* test, analysis of variance, and the test of a correlation. Permutation tests are for hypotheses about group differences or associations. As with parametric techniques, and as discussed in this chapter, the types of conclusions that can be drawn rely more on the design of the study than on the methods themselves.

The second group of methods involve the *analysis of categorical data*. Introductory categorical data analysis is often taught as part of an initial statistics sequence, though in this context it is rarely referred to as a nonparametric method. The categorical analysis test known as the chi-square test is a well-known technique that can be considered nonparametric because no assumptions are made about the form of the populations from which samples are selected.

Readers can obtain further information in one of many nonparametric textbooks, including those that are primarily written for the practitioner, such as Conover (1999), Sprent and Smeeton (2007), and Nussbaum (2015). These books often include chapters on categorical data analysis, but there are also entire textbooks devoted only to categorical data analysis (e.g., Agresti, 2013).

Table 26.1 Desiderata for Nonparametric Statistics.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. The use of nonparametric methods is justified.	M
2. The explanatory and response variables are clearly defined.	I
3. The level of measurement or categorization of the response variable is described.	I, M
4. The study design is clearly explicated.	M
5. Descriptive statistics are provided to convey variable effects or associations in the study.	R
6. The method of inference is matched to the characterizations of the explanatory and response variables.	M
7. The assumptions for valid inference are justified.	M
8. There should be a clear distinction between the use of exact calculations and large-sample approximations.	M
9. For rank-based methods, if ties are present, the methods used to address or adjust for ties are described.	M
10. Multiple comparisons are included when there are $K > 2$ levels of the explanatory variable.	M, R
11. Confidence intervals for the primary parameter of interest are included, when possible.	R
12. Conclusions refer to both descriptive and inferential findings and are consistent with the analysis outcomes.	R, D
13. Nonparametric extensions are considered for more complex designs.	M, R

* I = Introduction, M = Methods, R = Results, D = Discussion.

1. Use of Nonparametric Methods Is Justified

Reviewers who are not familiar with nonparametric statistics may be inclined to either ask the researcher to replace these methods with “more appropriate” parametric techniques or ask the researcher to provide a reason for using nonparametric methods. Nonparametric methods are frequently justified. Unfortunately, myths about nonparametric statistics have led some reviewers to consider these methods as “second class.” One of the most common myths is that a parametric procedure is more powerful than the parallel nonparametric procedure. Like most myths, this one has roots in a truth; namely, the correct parametric hypothesis test is the most powerful test *when the assumptions about the population of data are completely true*. The fact that “most powerful” is conditioned on adherence to some strict assumption is often forgotten, leading to the common belief that nonparametric methods are less powerful. Indeed, many studies (e.g. Blair & Higgins, 1980, 1985; Hodges & Lehmann, 1956; MacDonald, 1999; Randles, 1980) comparing the power of analogous parametric and nonparametric methods demonstrate that the nonparametric method is more powerful in many situations encountered in research studies, such as when the data are skewed. In fact, even if the assumptions of the parametric method are completely correct, the loss of power is slight with a well-chosen nonparametric method.

There are other myths that reviewers should be aware of and which are debunked in the corresponding sections later in this chapter. These include the myth that nonparametric results are only valid for samples and not populations, that nonparametric methods should only be used for nominal and ordinal data, that the hypotheses used with nonparametric methods are vague and not as informative as those of parametric methods, and that only hypotheses tests, and not confidence intervals, can be used with nonparametric methods.

Most researchers who commonly use nonparametric methods will realize that they need to defend this use to most reviewers. An appropriate defense is that the characteristics of the population are

unknown. Indeed, this can be argued as the case in most research studies. Some methodologists have written about the rarity of normality in research (e.g. Micceri, 1989). Another justification is that the research involves small sample sizes. This justification is used because with large sample sizes a researcher is more inclined to invoke the central limit theorem (CLT), even when the characteristics of the population are unknown. While it is true that the consequences of the CLT may result in correctly calculated Type I error rates when analyzing non-normal data, in such situations a nonparametric method can yield more power than the parametric counterpart. Thus, reviewers should consider arguments regarding the characteristics of the population as potentially valid, even in large sample situations. If the research does involve small samples, then the small-sample defense of the method is certainly appropriate because in this situation a parametric method might result in incorrect Type I error calculations.

Even though a method may be appropriate given what is known, or not known, about the population serving as the source of the data, reviewers also have to consider the knowledge base of readers. This may be one of the most compelling reasons to give pause to the inclusion of nonparametric methods in a manuscript. It may be widely accepted that readers understand how to read and interpret the results of one-way analysis of variance, yet less so that they can do the same with a Kruskal-Wallis analysis. The most appropriate way to address this issue is to require the researcher to explain the analytic outcomes and consequent conclusions in more detail than might be expected with better known techniques. A nonparametric researcher can even draw parallels with those well-known methods in explaining the outcomes. This would certainly be a more rational approach than to reject a method out of turn that may have more validity than better known methods. Researchers can simultaneously educate regarding both the context of the research and the methods used within that context.

2. Explanatory and Response Variables Are Clearly Defined

The selection of a specific nonparametric method depends on the nature of the explanatory and response variables. In this respect, it is no different from parametric statistical methods. The researcher should make clear what variables he or she views as the potential influences and what variables respond to these influences. Knowing these variables and the level of measurement or categorization of these variables (discussed in the next section) leads directly to the correct choice of an analytic method.

The variables should be explicitly identified in the introduction to the study. When the study involves an intervention (i.e., an experiment or a quasi-experiment), the explanatory variable assumes causal attribution and is typically referred to as an *independent variable*. In these cases, the researcher will observe changes in a response measure, often called a *dependent variable* in intervention studies, that appear to covary with the planned changes in the explanatory variable. When the study does not involve an intervention, the researcher is looking to determine if the level or category of a response variable changes when changes occur in the explanatory variable, with the recognition that this response variation is potentially due to variables other than the explanatory variable. (Note that the researcher may still use the terms *independent* and *dependent* for the explanatory and response variables, even though there is no purposeful manipulation of the explanatory variable.)

As with parametric statistics, there can be more than one explanatory variable as well as more than one response variable. Multiple explanatory variables are additional factors or covariates. This chapter focuses on those methods with a single explanatory variable and either a single response measure or repeated response measurements, as this reflects the majority of nonparametric methods reviewers encounter in social science research. The inclusion of multiple factors and covariates offers many advantages, and nonparametric methods are available for such designs. Unfortunately,

these are given short shrift in social science education so that relatively few researchers use these techniques. Reviewers need to be aware that such methods exist so that if these are encountered in the research that they are not dismissed out of turn. They may, in fact, be cutting edge.

3. Level of Measurement or Categorization of the Response Variable Is Described

It is always important to identify the level of measurement used for the variables. In fact, if careful attention was paid to this topic, more researchers might be inclined to select a nonparametric method over a parametric method. It is not uncommon to find researchers using methods that assume quantitative measures when, in fact, the data are ordinal (i.e., ordered categorical). That is not to say that the level of measurement strictly defines the appropriate use of parametric or nonparametric methods. Nonparametric methods can be used with any type of data, though many parametric methods assume both quantitative response measures and a particular form to the population that gave rise to these data. This suggests that there are times when reviewers should inquire as to why a parametric method is being used instead of a nonparametric method, rather than the other way around.

Data can always assume a “lower” (or coarser) level of measurement, but the converse is not true. For example, quantitative data can be placed into ranks or categories, but ordinal data cannot be transformed into quantitative data without further measurement. This is why the many rank-based nonparametric methods can be used for both ordinal and quantitative data. This suggests two considerations that reviewers should tend to when reviewing the level of measurement. First, are the data truly quantitative? A survey made up of Likert items can indeed yield quantitative data if rigorous scaling techniques are used during the instrument development process. If that is not the case, then it is likely that, at best, the data are ordinal. The fact that the total score of a survey is a numeric value does not justify the assumption of quantitative data. Numbers might only reflect order and not quantity. The second consideration is about the assumptions required for the validity of a statistical method. Even when the data are quantitative, there may be reason to suspect that one or more assumptions are probably not tenable. In such cases it is often true that a nonparametric method, which has fewer assumptions than the parametric counterpart, is a valid choice because the assumptions hold for this method.

4. Study Design Is Clearly Explicated

As with all inference, the method of sampling influences the kind of conclusions that can be drawn from nonparametric analysis. Inference to a particular population of individuals with a specified level of confidence requires that the researcher randomly selected individuals from that population. This is seldom the case in social research. Researchers should be clear in their conclusions that inferences made without random sampling only apply to an unspecified and unknown population that is represented by the study sample. This can still be meaningful in the quest for new knowledge, but it is quite different than providing a statement about specific population parameters with a supposed level of confidence.

Nonparametric methods can be a good choice for causal inference about variable relations in the sample. This does not require random sampling but *does* require randomization, the random assignment of individuals to conditions. When a study includes such randomization and the (typical minimal) assumptions are correct, then nonparametric analysis, which itself involves a randomization process known as a *randomization test*, yields appropriate statistical inference and correct confidence levels about the causal relations among the explanatory and response variables. Reviewers need to be on guard for a researcher who uses the randomized

nature of the analysis as a justification for making causal claims. If the study design did not include randomization of participants, then the use of a randomization test in the analysis does not yield causal results.

Computer algorithms for nonparametric analysis will yield results regardless of how the researcher obtained the data, so it is up to that researcher (with a reviewer check) to make certain that the claims match the method of sampling and the study design. In short, random sampling can lead to inference about existing differences. Randomization can lead to inference about causal relations. In the rare case when these are both present, causal relations can be inferred as a potential for a larger, and specified, population. In the more common case where neither random sampling nor randomization is possible, then neither parametric or nonparametric analysis leads to correct probability statements about parameters or causes, but serve only as a type of so-called *calibration inference* (Draper, 1995) to moderate the human inclination to detect patterns in just about any observation.

5. Descriptive Statistics Provided to Convey Effects or Associations

When discussing results, it can be tempting for a researcher to focus on *p* values, confidence intervals, and the tools of inference as the “sophisticated” findings in a study. Curb such enthusiasm. It is important for researchers to report on sample outcomes. When the center of the data distribution, or differences in the centers of multiple distributions, is the primary focus, with nonparametric statistics the primary central tendency statistic is usually the median. The nonparametric description of variability typically focuses on percentiles and quartiles. Keep in mind that many nonparametric statistics are appropriate for non-normal distributions, including skewed distributions where the median is the more appropriate indicator of center than the mean. The standard deviation relies on deviations about the mean, so this is also problematic in the presence of skew. When the data are near symmetrical, the mean and median will be similar, so that either (or both) can be used as an indicator of center.

Nonparametric methods that rely on ranks require a transformation when the data are quantitative. This simple transformation, from quantities to the ranks of these quantities, results in descriptive outcomes that are based on these ranks, such as means of ranks. As with the original data, these rank-based statistics can and should be used to discuss group differences, rather than relying on a *p* value to accomplish this task. Readers are unlikely to be as familiar with the meaning of differences in mean ranks as they are differences in means, so researchers should discuss this meaning. Quantitative mean differences can place the center of one sample above that of another by a specified amount, whereas mean rank differences refer to the individuals in one sample being more likely to be above those in the other sample on the response measure. This is a correct interpretation even when the data are ordinal, which cannot be said about mean differences. The use of mean differences is inappropriate when the data are ordinal.

The nonparametric measures of association when the data are category frequencies include differences in proportions, Kendall’s tau, and Cramer’s *V*. These may be used to discuss the strength of relation among the explanatory and response variables. Proportional differences typically provide the most intuitive understanding of the size of the effect. Associational measures, such as Cramer’s *V*, should be carefully explained. *V* has a range from 0 to 1, and as such it is like quantitative associational measures, such as Pearson’s correlation coefficient. The problem, however, is that even small non-zero values of *V* can signify substantial association. That is because *V* reflects association as a proportion of the maximum possible association. In many contexts where frequencies are analyzed, small differences in the relative frequency of occurrences in these categories are sufficient to suggest meaningful association.

6. Method of Inference Matched to Explanatory and Response Variables

There are many more nonparametric methods available than what will fit comfortably in a single chapter in this book, yet the list provided in this section captures most of the methods that a reviewer is likely to encounter in the social science literature. Researchers should select a method by identifying (1) the number of explanatory and response variables, (2) the level of measurement or type of categorization used to operationally define these variables, and (3) the number of categories used for this definition when at least one of the variables is categorical. This is one reason it is so important for the researcher to not only identify the variables, but also to discuss how these variables are defined in the study. The nonparametric method of inference, whether it be a test of a hypothesis or the construction of a confidence interval for a parameter, must be a correct match to the properties of the variables.

(a) *Sign test.* The sign test is a hypothesis test to determine whether there is a greater proportion of individuals in one of two dichotomous categories. The most obvious use for this test is in a matched pair situation, with one member of each pair being in a treatment condition and the other member in the control condition. The question of interest is whether there is sufficient evidence to demonstrate that the treatment is effective. The sign test helps accomplish this task by allowing the researcher to test, with a specified level of confidence, whether members of the treatment condition were more successful than their control match. This test nicely illustrates the principle that quantitative data can be categorized, if need be. The method for ascertaining that the member of the matched pair who receives the treatment is “more successful” can be through dichotomous categorization (e.g. completed the task vs. not completing the task) or through quantitative difference scores. Thus, the sign test can be used instead of a matched-pair *t* test, with only the sign of the difference scores needed for the analysis. Researchers will be correctly using the sign test if they are focusing on a single dichotomous variable, or if they create a dichotomy from a single quantitative variable, such as difference scores.

The sign test is a special case of the binomial test that is presented in many introductory statistics texts (both parametric and nonparametric). In fact, a benefit of using the more general binomial test instead of the sign test is that a researcher can construct a confidence interval for the proportion of individuals (or matched pairs) that will exhibit a desired outcome, such as treatment benefit. Reviewers should bring this to the attention of a researcher who simply presents a *p* value for the sign test rather than taking the additional step of creating a confidence interval for proportion success.

It is also important to note that the one-sample Wilcoxon test (see below) is a competitor to both the sign test and the matched pair *t* test when the data are quantitative. In fact, the Wilcoxon test should frequently be selected instead of the sign test. The sign test is only a more powerful test in those situations when the tails of the population distribution are very heavy, such as what we might expect when we collect opinion data regarding a polarizing issue.

(b) *Goodness-of-fit tests.* The sign test is for determining if there are a higher proportion of individuals in one of two categories. The goodness-of-fit tests extend this to more than two categories. In order to be able to conduct such a test, the researcher must have a valid reason for specifying specific proportions for each of the categories. For example, suppose that the proportion of individuals fitting into one of four personality types is known from multiple studies. These proportions could be hypothesized for a sub-population to determine if this population is unique with regard to this personality distribution. Note that the goodness-of-fit test is conducted using a single sample and a single categorical response variable with *K* mutually exclusive possible categories, and not with *K* samples. There are several possible goodness-of-fit tests, including the Karl Pearson goodness-of-fit test, the Kolmogorov goodness-of-fit test, and the Lilliefors test, which is a special case of goodness-of-fit in that it tests whether the data come from a normal distribution.

(c) *One-sample Wilcoxon test.* The one-sample Wilcoxon test is sometimes referred to as the *one-sample Wilcoxon sign test*, but is different from the sign test described above. It is conducted with a single sample of either ordinal or quantitative data, and as such is a competitor of the one-sample *t* test but is available in a broader set of research situations. This test relies on ranks. It serves as a good example of why nonparametric tests should not be routinely rejected by reviewers. The Wilcoxon test has fewer distributional assumptions than the *t* test, is appropriate when the data are ordinal (unlike the *t* test), often is more powerful than the *t* test, and is only slightly less powerful than the *t* test when the assumptions for the *t* test are valid, including that of a normal distribution.

Many researchers are unaware that when the data are quantitative, the one-sample Wilcoxon method can be used to construct a confidence interval for the median. These researchers are prone to provide only descriptive statistics and a *p* value. Confidence intervals are more informative than *p* values, so it is important to encourage researchers to learn how to construct the corresponding confidence interval when using the one-sample Wilcoxon test.

(d) *Irwin–Fisher test.* This is often referred to as the *Fisher exact test*, though Fisher's exact method extends to many nonparametric tests, and not just this one. The Irwin–Fisher test is used to compare the proportion of successes in two groups, where “success” can be defined as any specific outcome (e.g. passing a test, obtaining a job). For this so-called “exact test” to provide exact and correct *p* values, the total number of successes must be known in advance. For example, suppose that judges rate tryouts for a district orchestra with the idea of filling ten vacant positions. It may be of interest as to whether there is bias toward time of day of the tryout (morning or afternoon), so the proportion of successes for the morning tryouts is compared to the proportion for the afternoon tryouts. The number of tryouts is known in advance, as is the total number of successes. What is not known is how many of these successes will come from each group.

In most situations, the total number of successes will probably not be known in advance, in which case researchers can use lesser known methods, such as Barnard's exact test or the product binomial test. For larger samples, a chi-square test (see below) can be used as a good approximation in both the situation when the total number of successes is known in advance, as well as when these are not known in advance. The two-sample median test is a useful special case of the Irwin–Fisher test. This test is for ascertaining whether two population medians differ. In this case, “success” is above (or below) the median, and since the median splits the data in half, the number of successes are known in advance, which is why the Irwin–Fisher test is appropriate. For the situations discussed above, when sample sizes are at least moderately large (i.e., five or more observations in each of the four cells of the table), the researcher should provide an approximate confidence interval for the difference in proportions.

(e) *Mann–Whitney test.* This is also called the *two-sample Wilcoxon test*. (The two tests were created independent of one another and use different procedures, but were later shown to be equivalent tests.) The test is for comparing two populations with ordinal or quantitative response data. As such, it is a competitor to the two-sample *t* test, though it has broader use because the *t* test should not be used with ordinal data. As with the comparison of the one-sample Wilcoxon test to the one-sample *t* test, the Mann–Whitney test can be more powerful than the *t* test in many practical research situations and is almost as powerful as the *t* test even with normally distributed data, which is an assumption of the *t* test. A confidence interval for the median difference score is a valuable statistic that researchers should provide. Note that the median difference is *not* the same as the difference in medians, so there is benefit to researchers explaining this statistic to their audience.

(f) *Kruskal–Wallis test.* The Kruskal–Wallis test is a nonparametric alternative to one-way between-subjects analysis of variance. The explanatory variable is categorical, with more than two mutually exclusive categories, and the response data are ordinal or quantitative. As with ANOVA, the Kruskal–Wallis test is for comparing more than two populations. The criteria for selecting the

Kruskal–Wallis test or one-way ANOVA parallels those for the Mann–Whitney test and two-sample *t* test. The Kruskal–Wallis test has fewer assumptions, can be used with ordinal data, is often more powerful than ANOVA, yet is almost as powerful as ANOVA even with normally distributed data. Whereas the Type I error rates for one-way ANOVA may not be held at the desired level, the exact Kruskal–Wallis test does control the Type I error rate. A historical reason for not using the Kruskal–Wallis test is that the critical values were only available for small sample sizes (i.e., no more than five observations in each group). There are now tables of exact Kruskal–Wallis critical values for total sample sizes up to 105 (Meyer & Seaman, 2014), which covers many practical research situations, so that researchers should no longer shun the Kruskal–Wallis test or provide large-sample approximations in settings where exact critical values are available.

(g) *Friedman test.* The Friedman test is an extension of the sign test with individuals ranked on more than two response categories. It is often considered the nonparametric alternative to one-way within-subjects analysis of variance, but this is not always the case because with the Friedman test ranking is conducted separately for each individual, without regard to the relative position of that individual's scores. Similar to within-subjects ANOVA, however, the test can be used with either repeated measurements or blocks (i.e., individuals who have been matched on a correlated extraneous variable). The hypothesis of interest is whether the median ranks differ across the repeated measurements or across the correlated groups. This is appropriate if the researcher is interested only in relative order of individuals across conditions, but often researchers are also interested in the size of these differences. If this is the case, the researcher can be encouraged to consider lesser-known alternatives, such as the Hodges–Lehmann aligned ranks procedure (Hodges & Lehmann, 1962) that involves ranking across all scores, rather than separately for individuals. A nonparametric within-subjects procedure should be used instead of the usual repeated measures ANOVA if the data are ordinal, but should also be used when the data are quantitative and there is a violation of the assumptions necessary for repeated measures ANOVA. When the response data are dichotomous (e.g. “yes” or “no”), the Friedman test can be used but in this specific case is known as *Cochran's Q test*.

(h) *McNemar's test.* McNemar's test is used to compare two proportions from the same individuals when the response data are dichotomous. It is a one-sample test in which the individuals in the sample are placed into one of two categories, and then are placed again into one of the same two categories. For example, the test can be used to compare the proportion of awareness about a topic before and after a public service campaign. There are extensions to more than two categories (Stuart's test) as well as to more than two measurements (Cochran's Q test). With moderate to large sample sizes, confidence intervals can be calculated for both McNemar and Stuart proportional differences, so reviewers should insist on these when the sample size makes it tenable.

(i) *Spearman's correlation and Kendall's tau.* Both Spearman's correlation and Kendall's tau are rank-based measures of association. As such, they can be used with either ordinal or quantitative data and both assess the relation between two variables. The variables can be of the same or different data types, so long as the data types are ordinal or quantitative. In either case, the data are transformed to ranks.

Spearman's correlation coefficient is calculated by applying Pearson's correlation to ranks. Because it uses ranks, rather than the original scores, it is a measure of monotonic association rather than linear association. That is, it indexes the degree to which two variables increase (or decrease) together without regard for the magnitude of change. An ever-increasing curvilinear relation between two variables will result in a Spearman correlation of 1, whereas the Pearson correlation will be something less than 1 due to non-linear association. Although Pearson's correlation is the most widely used measure of association, in many cases it is probably Spearman's correlation that is of interest. Many researchers are interested in knowing if increases in one variable result in increases in a second variable and are not strictly interested that these changes take a linear form.

Kendall's tau is calculated by considering whether the ranks from pairs of individuals are concordant (i.e. both increase or both decrease) or discordant (i.e. move in opposite directions). Specifically, it is the difference in the proportion of concordant and discordant pairs out of all possible pairs. As with Pearson's correlation, both correlational coefficients range from -1 to 1 and thus it should not be surprising that these two nonparametric correlation indices are themselves correlated. Confidence intervals are also available for both and should be used instead of, or in addition to, the p value of the test. In the case of correlations, the p value is typically based on a hypothesis of no correlation, which is seldom very informative because what is of interest is how large of an association exists among the variables, rather than whether there is any association at all.

(j) *Tests of homogeneity and association.* When most researchers hear of the chi-square test, they probably think of a test of association between two categorical variables. That is indeed the most widely known use of the chi-square test, though it is also a generalization that does not consider the potential methods of sampling. There are three possible sampling methods that can be associated with the two-way contingency (sometimes called *crosstabs*) table that we think about as leading to inference based on the chi-square statistic. The first, and rarest, of these is when a specified number of individuals are selected from each of K populations and categorized on a response variable such that the total number in each category of this response variable is known in advance, but it is unknown how many units will be in each category within each of the K samples. In this case, both margins of the contingency table are known prior to data collection. An example of this is the median test, discussed above. For this test, we know that half of the total sample will be categorized above the median and half below, and we know the sample sizes for the K samples, but we don't know the number in each response category within each sample. What is of interest in this case is whether the proportions in each response category is the same in each of the populations.

The focus on proportional similarity (or difference) is also true in the far more common situation when K samples are selected and compared on response categories, but neither the proportion of responses in each category within the samples or total for all samples is known in advance. In this case, only one margin of the contingency table is known in advance because the sizes of the samples that will be selected are known. In this situation, the use of a chi-square statistic for an inferential test of this contingency table is a test of homogeneity of proportions across the K populations for each of the response categories. By contrast, a test of association of two categorical variables is conducted using a chi-square statistic when a single sample of individuals is categorized on each of two response variables. Once again the data are displayed in a two-dimensional contingency table, but in this case the only frequency known in advance is the total number in the sample. Neither of the margins of the contingency table are known in advance.

This difference in the number of margins that are fixed in advance (zero, one or two margins) requires different types of nonparametric inference if the researcher conducts an exact test. Such tests are seldom chosen in social science research, but it would be well for reviewers to understand that when an exact test is used in this situation, the researcher should clarify the choice of the test based on the number of margins that are fixed in advance in the contingency table. The far more common situation is for the researcher to conduct a large-sample test. The most common rule of thumb used for determining if a large-sample test is appropriate is that the expected cell count is at least 5 in at least 80% of the cells in the contingency table. In most research situations, this is easy to achieve, so the large-sample test is often appropriate. Serendipitously, in all three of the marginal situations, the chi-square statistic approaches a chi-square distribution so that the common chi-square test is appropriate. This is likely the reason that the differences in the types of tests of frequencies with two categorical variables is often not even considered. Reviewers should make certain that researchers do indeed have an adequately sized sample (or samples) to at least minimally meet the rule-of-thumb criterion.

For these large-sample tests, proportional differences among K populations or association of two categorical variables is shown when the chi-square statistic is large enough to yield a small p value. For the test of homogeneity, sample proportions can be used to estimate proportional differences. This should be used as an indicator of effect, rather than the p value. Confidence intervals can also be constructed for these proportional differences. The descriptive estimate of association is Cramer's V statistic, though in practice this is difficult to interpret because it is based on the maximum possible association that can be achieved with the observed marginal frequencies. It is often the case that Cramer's V is small, even when there is meaningful association between the two categorical variables. Researchers should not interpret V using the same rules-of-thumb that they might use for Pearson's correlation coefficient. With contingency tables, it is often easiest to interpret strength of association by comparing proportions. Even though this is a technique for the test of homogeneity rather than that of association, the fact that these large-sample methods converge provides some justification for looking at proportional differences even for the model of association.

7. Assumptions for Valid Inference Are Justified

Nonparametric methods are often referred to as *assumption-reduced* methods. This is not the same as *assumption-free*. As with all methods of statistical inference, certain conditions must be met for the inference to be valid, and researchers should pay attention to these conditions. Fortunately, a nonparametric counterpart to a method of parametric inference often dispenses with the conditions that are most difficult to achieve. Another valuable characteristic of some nonparametric methods is that when an assumption cannot be verified, the method may still lead to valid inference, though a different and more general inference than what could be made when the assumption is valid. Thus, reviewers should check that the conditions for inference have been discussed and that the inferences made are consistent with those conditions that could be verified or reasonably assumed.

The lettering that follows in this section corresponds to that of the previous section, with the same methods discussed. In this section the focus is on the conditions for valid inference for each nonparametric method.

(a) *Sign test.* The only assumptions necessary for valid inference with the sign test are that the individuals are independent of one another and that there is enough evidence to determine which of two response categories best represent each study participant. In the matched pair situation, for example when matching up a member of a control condition with one in a treatment condition, the members of each pair are not independent, but are matched on an extraneous variable. In this case, it is the pairs that must be independent of one another.

Sometimes the continuity of a response measure is listed as an assumption of the matched-pair sign test, which may be confusing given that the response is one of two categories. This continuity assumption is simply a way of stating that the researcher must be able to determine which member of the matched pair is most successful. If the researcher has provided convincing evidence that this comparison is valid, then independence of pairs is the only requirement for valid inference.

Some textbooks discuss using normal distribution methods to provide a large-sample approximation for the sign test, so reviewers may encounter researchers who employ this technique. These reviewers should be sent back to the drawing board because such a large-sample approximation is unnecessary. With appropriate software, an exact sign test can be easily calculated for any size sample and the exact test is preferred to approximations.

(b) *Goodness-of-fit tests.* These tests have the same requirements as the sign test: independence of observations and the ability to correctly classify these observations into one of K mutually exclusive categories. Note the term "mutually exclusive." This means that each participant should fit into one and only one category. A misuse of this test is when individuals are permitted to respond to multiple

categories, and then the frequency in each of these categories is incremented. The goodness-of-fit test should not be used in this situation. A simple check a reviewer can conduct is to determine if the sum of the frequencies in the one-dimensional contingency table is equal to the sample size. It should be.

(c) *One-sample Wilcoxon test.* The one-sample Wilcoxon test is a nonparametric competitor to the one-sample *t* test. The *t* test is especially important in the matched-pair situation where the intent is to determine if the mean of one population (e.g., the population that can receive an intervention) is greater than the mean of a second population (e.g., the population without intervention). In this case, the assumptions of the *t* test are that (i) participants are measured on a quantitative response measure, (ii) the pairs are independent of one another, and (iii) the difference scores are normally distributed. Note that this third assumption will be true if the distributions of scores for both the experimental and control populations are normal.

The assumptions for the one-sample Wilcoxon test parallel those of the *t* test with the assumption of normality reduced to an assumption that the distributions either have the same shape or are both symmetrical. In these cases, the null hypothesis can be for equal medians or equal means, as the difference will be the same for each. An assumption of similar shape would be tenable if the intervention has a similar effect to individuals across the entire range. This may sound implausible, but not only is this the same expectation as for the one-sample *t* test, the *t* test also includes an assumption of normality.

If the researcher cannot make a reasonable case that the populations have the same shape or are symmetrical, the Wilcoxon test can still be conducted, but the hypothesis changes. Instead of testing equality of medians or means, the researcher can now test whether there is a greater chance that the member of the treatment will have a higher (or lower) score than there is that the matched member of the control condition will have a higher score. Reviewers, and the public at large, are likely not to be as familiar with this hypothesis, yet it often addresses the primary research question: Does the treatment work? Not knowing much about population distributions leads to a loss of specificity, yet this is probably a more realistic view of many research studies. This less-specific, yet still important, hypothesis should also be used when the data are ordinal. Ordinal data is often collected in social research, and in these cases reviewers should insist on the one-sample Wilcoxon test, rather than the *t* test.

(d) *Irwin–Fisher test.* This method is only valid when both margins of a contingency table are known in advance and each observation is placed into one of two mutually exclusive categories. Aside from that, the only assumptions are that the two samples are independent of one another and observations within each sample are also independent of each other.

(e) *Mann–Whitney test.* The assumptions of the Mann–Whitney (or two-sample Wilcoxon) test are that the two samples are independent of each other, observations are independent within each sample, and the populations from which the samples were drawn have the same shape. For the analogous two-sample *t* test, this third assumption is that the distributions have the same shape and this shape is normal. Thus, as with the one-sample Wilcoxon test, the primary difference in assumptions is a relaxation of the distributional form from normal distributions to same-shaped distributions. Researchers should make the case that any intervention has a uniform effect across the range of the distribution so that it shifts the location of the distribution without altering distribution form. If the researcher cannot reasonably make this case, the Mann–Whitney test can still be used, but with a different hypothesis, namely, that higher (or lower) scores are more likely to come from the treatment population than from the control population. As with the one-sample Wilcoxon, the less specific hypothesis is likely to be consistent with the heart of the research question and should be used when there is uncertainty about distributional shape equality or when the data are ordinal.

(f) *Kruskal–Wallis test.* A valid Kruskal–Wallis test depends on the K samples being independent of one another and the observations within these samples also being independent of each other. To test the equality of medians (or means) in the underlying populations, the population distributions must have the same shape across the K populations. Fortunately, the Kruskal–Wallis test is more sensitive to differences in location than it is to other variations in population shape (Srisukho & Marascuilo, 1974). A researcher does not necessarily need to make a strong argument that a treatment shifts a population distribution without changing the shape of the distribution, though it is important to provide evidence that location is shifted along with, or even better, instead of, other aspects of distributional form.

As with the Mann–Whitney test, ordinal data can be used, but this requires a change in the hypothesis. With these data, the hypothesis is that the median ranks are the same across the K conditions. This parallels the hypothesis of the Mann–Whitney test for ordinal scores. Namely, higher median ranks reflect that the probability is greater of obtaining a higher score (ordinal or quantitative) from these conditions than it is for obtaining higher scores from other conditions with lower median ranks.

(g) *Friedman test.* The Friedman test assumes that individuals are measured multiple times using either a quantitative or ordinal measure. Though these measures are correlated, the individuals themselves must be independent of one another. There is no assumption of normality. As discussed above, before using the Friedman test, the researcher should clarify that what is of import is only that some conditions consistently yield higher scores or ranks regardless of the magnitude of individual differences.

(h) *McNemar’s test.* McNemar’s test is appropriate if participants are independent of each other and each participant is categorized twice using dichotomous categories. This should not be confused with a test of two independent proportions. In the latter case, the proportions are compared from two independent samples. Conversely, the test of two independent proportions should not be used if both proportions are obtained from the same set of individuals.

(i) *Spearman’s correlation and Kendall’s tau.* The only assumption for valid inference for nonparametric measures of association is that individuals are measured independently of each other. There is no assumption regarding the distributional form of these measures. As with all statistics, parametric or nonparametric, the validity of inference to a larger population depends on a sample that is representative of this population. Reviewers should ascertain that researchers are not claiming an association in a population based on a sample that is not representative of that population (e.g., a random sample).

(j) *Tests of homogeneity and association.* Whether testing for homogeneity of the proportions of responses in K samples, or assessing whether there is an association between two categorical variables, individuals must be categorized independently of each other and each participant must contribute to only one cell frequency in the contingency table. An all-too-frequent misuse of the chi-square statistic with contingency tables occurs when participants are measured more than once and included in the counts in multiple cells. As with the goodness-of-fit test, a simple check that reviewers can make is to ascertain that the total of all cell frequencies is equal to the size of the sample.

8. Exact Calculations vs. Large-Sample Approximations

Each nonparametric method can be used to calculate inferential statistics one of two ways: using an exact method or using a large-sample approximation. Exact methods are called “exact” because the methods yield precise probabilities for p values and confidence intervals when making inferences based on the sample data. For these probabilities to truly be exact, the required assumptions

for inference must be correct and the design of the study must be consistent with the inference. The small number of assumptions needed for most nonparametric methods make it feasible that these assumptions are valid, especially if the researcher chooses to test a broader hypothesis and only rely on easily verified assumptions, such as the independence of the observations. It is usually more difficult to match the study design to the desired type of inference. For example, constructing a confidence interval for the median of a population using exact methods will not provide an exact confidence interval if the data come from a non-random sample. In this case, the term “exact” is a misnomer. This is not a nonparametric problem, but an inference problem that applies to both nonparametric and parametric statistics alike. A confidence interval for a mean constructed using t -test methods is an exact confidence interval if the assumptions are truly valid (including an assumption of a normally distributed population of scores) and the sample was obtained using random sampling methods in which each member of the population was accessible and equally likely to be chosen for the sample. Such is rarely true in social research. This does not invalidate the methods, but reviewers should be aware that a researcher who refers to exact methods cannot claim to be providing exact probabilities in most situations.

Large-sample approximations are used in nonparametric statistics to approximate the results that would be obtained if exact methods could be used. It used to be the case that large-sample approximations were needed even when samples were relatively small. This is because many computations are required to calculate exact values. More and more, however, fast computers and modern algorithms can yield exact methods even with moderate to large samples. Reviewers who encounter researchers using large-sample approximations with nonparametric methods and relatively small sample sizes should call into question whether such an approximation is necessary. It may be the case that the researcher is using an older table of critical values or software that has not incorporated newer algorithms. Newer tables of critical values are much more expansive, and specialized algorithms for calculating exact probabilities are more broadly available, negating the need for both critical value tables and approximations.

It is still the case, however, that sometimes large-sample approximations are needed, especially for confidence interval construction. In fact, researchers who provide an exact p value and no confidence interval probably are ignoring, or are unaware of, large-sample approximations that can be used to construct informative confidence intervals. Reviewers should insist on these intervals. Even if a large-sample formula is not available for the interval, for most nonparametric methods there is software available to estimate the interval using Monte Carlo methods.

Nonparametric exact methods are based on discrete data. With a sample of n integers, only certain p values are possible when using exact methods. Given that large-sample approximations are based on continuous distributions, such as the normal distribution, yet are being used to estimate exact values, the large-sample approximation should include what is known as a “correction for continuity.” This correction slightly increases p values and widens confidence intervals to reflect the fact that a continuous distribution was used to calculate a discrete value. The larger the sample size, the smaller this correction, so that it is only needed with small to moderate sample sizes. In these cases, however, a thorough researcher should specify that they used a large-sample approximation with a correction for continuity.

9. Methods to Adjust for Ties Are Described

Nonparametric methods, particularly exact methods, assume that decisions can be made regarding the relative order of two observations. This is the reason that a continuous measure of the underlying construct is often listed as an assumption. In practice, this means that there will be no ties among observations in the data. This “no-ties rule” does not appear with parametric

procedures because the actual scores are aggregated with various functions, such as those that determine the mean and standard deviation. In contrast, many nonparametric methods rely on ranks, so it is assumed that the data can be completely ranked, thus necessitating the ordering of all pairs of observations.

Reviewers of manuscripts using nonparametric statistics should not only be aware of the “ties” requirement, but also a reasonable approach to addressing the tied observations that often occur, the requirement notwithstanding. There are three traditional methods for dealing with ties. One method is to discard tied observations. This has the disadvantages associated with discarding data in any study, including the obvious disadvantage of reducing power by reducing the size of the sample. A second method is to randomly break each tie. This is a theoretically reasonable solution given that inference is based on long-term probabilities, yet it has the practical disadvantage that two researchers who obtain the exact same set of data and apply the same methods to these data can reach different conclusions regarding inference. For these reasons, the third method has gained the most acceptance; namely, to assign midranks (i.e., the average of the ranks) to the tied observations. This is not a perfect solution because the so-called exact methods applied to data that includes midranks yield approximate p values and confidence intervals. This approximation may be quite different than the exact values, especially with smaller samples. Fortunately, there are adjustments for tied observations that bring the approximations close to the correct probabilities. An astute reviewer will ascertain that researchers who encounter ties in their data have used midranks and have employed an adjustment for ties.

There are two more noteworthy points regarding tied observations. First, for a few nonparametric methods, the use of midranks can create problems. For example, in the median test, every observation is classified as being above or below the median. How should a researcher proceed if a midrank corresponds to the median? There are two acceptable solutions. One is to classify both observations either above or below the median so that a hypothesis test becomes more conservative. That is, the researcher makes it more difficult to reject the null hypothesis of median equality of the control and treatment conditions. The obvious downside to this is the loss of power, but the advantage is that Type I errors are controlled below the acceptable level. The second method is to place one observation above the median and one below. This is preferred by researchers (and reviewers) who are not squeamish about slightly excessive Type I error rates and consider this preferable to a forced decrease in power.

The final point is about a lesser known approach to nonparametric methods that removes the need for separate test statistic formulas when there are ties. This is best explained with an example. Consider the Kruskal–Wallis test for the equality of K population distributions. The traditional method of calculating the test statistic uses a so-called “computational formula.” In fact, there are two such formulas: one to use without ties and one to use when adjusting for ties. Like most such formulas, these were created to provide ease of hand computation in an era that preceded readily accessible computer power. With computers, the Kruskal–Wallis H statistic can be calculated from sums-of-squares. With this method, the H statistic is correct regardless of whether there are ties in the data. If there are ties in the data and this H is compared to a table of critical values, it is not a perfectly exact test, yet the correction for ties is built in and the result is the same as if the researcher had used the ties-correction formula.

Even more cutting edge, a researcher can create all permutations of the data, calculate the test statistic for each permutation, and then ascertain whether the observed statistic falls in the tails of the resulting distribution. This provides an exact test even when there are ties. This method is most feasible with smaller sample sizes, but for larger sample sizes researchers could take a random sample of all possible permutations, thus using a Monte Carlo approximation to exact values that again works just as well with or without tied observations.

The above considerations will rarely need to be considered as these methods are not widely known, yet it is valuable for a reviewer to be aware of the possibilities. First, a reviewer encountering such practice should applaud it, rather than reject it. Second, we can hope that the training of practitioners in the use of nonparametric methods will reach the point such that modern computing power is utilized to calculate more precise inferential statistics regardless of whether there are tied observations in the data.

10. Multiple Comparisons for $K > 2$

Omnibus hypotheses are useful for exploratory analysis, but testing these hypotheses seldom answers the specific questions of interest in a research study. Parametric methods have long been used to conduct specific multiple comparisons that control Type I error for a family of such comparisons, yet social researchers who employ nonparametric methods often are unaware that similar planned or post hoc methods are also available with reduced assumptions. Indeed, some researchers may favor a parametric method because they believe this is the only way to construct confidence intervals that address questions of interest, regardless of whether the parametric assumptions are tenable.

Two common settings will exemplify how reviewers can guide researchers to provide specific comparisons. In a simple omnibus design, K groups will be compared on either a categorical or quantitative response variable. When the response variable results in classification of the observations into unordered categories, this is the model of homogeneity described above. If the response variable is ordinal or quantitative, the researcher could use a Kruskal–Wallis test. In either case, providing an omnibus test statistic and stopping there fails to address the question about differences among the K groups. For both situations, there are methods for either foregoing the omnibus test entirely and examining specific differences, such as pairwise comparisons, or first testing the broad omnibus hypothesis and then following this with more specific comparisons. This parallels the more commonly known parametric situation in which the one-way analysis of variance is precluded or followed by comparisons based on the t distribution.

The details for analysis are beyond this chapter and can be found in a good nonparametric textbook, but a simple description should be sufficient to help reviewers guide researchers to these details. For the model of homogeneity, what is often of interest is how much of a difference there is in the proportion of observations that fall in one category for one of the K groups and the proportion that fall in that same category for a different one of the K groups. This proportional difference is easily described for the sample, but an additional confidence interval extends this to inference beyond the sample data. Additional confidence intervals can be constructed to compare different groups or different categories. The researcher should use a method for controlling Type I errors across the set of intervals. In this context, error control could include a Bonferroni procedure or a procedure analogous to Scheffé's method with analysis of variance comparisons.

For a quantitative response variable when the Kruskal–Wallis method is the appropriate omnibus method, there are two ways that researchers can calculate confidence intervals. One method is to compare specific groups using the mean ranks obtained for the Kruskal–Wallis test. These comparisons are based on jointly ranking across groups, and as such can result in a loss of power in many group configurations. Additionally, the confidence intervals are difficult to interpret. A better solution is to create Mann–Whitney confidence intervals for pairs of conditions, re-ranking for each pair and using Bonferroni Type I error control. This results in easy-to-interpret confidence intervals that provide specific inference about conditions.

11. Confidence Intervals

Confidence intervals go beyond the two multiple comparison settings discussed in the previous section. Indeed, confidence intervals are available using nonparametric methods in most contexts. This fact is not well known, even though it is likely that most researchers now know to use confidence intervals instead of, or in addition to, p values when such intervals are available. Thus, it behooves reviewers to push researchers to construct confidence intervals even when using nonparametric methods. These intervals are commonly for the difference in proportions or medians, though they can also be for the difference in means in the cases when symmetrical or identical distributions are assumed across the populations of interest.

As with other nonparametric analyses, confidence intervals can be constructed using one of two flavors: exact methods or large-sample methods. Confidence intervals constructed using exact methods rely on the inversion method of construction. In brief, with this method specific hypotheses, such as the size of the difference in medians, are tested, varying the hypothesis for repeated testing. The confidence interval consists of all retained (i.e., not rejected) hypotheses. Many, but not all, nonparametric methods can be used to construct exact confidence intervals using this procedure, but some methods require large-sample approximations that are based on mathematical distributions, such as the normal distribution. The important point is that almost always such intervals can be constructed. Reviewers noting researchers providing p values without providing confidence intervals should push to see intervals, and only relinquish when it can be shown that such intervals are not available. This latter situation is rare.

12. Conclusions Refer to Descriptive and Inferential Findings, and Are Consistent with Analysis Outcomes

As with research that includes parametric analysis, conclusions drawn from nonparametric analysis should match the analytic results. Perhaps the greatest obstacle some researchers face in stating conclusions drawn from nonparametric analysis is that of correctly interpreting analysis using score transformations, such as ranks. Part of the difficulty is that the same analysis of the data can result in various types of conclusions, depending on the type of data and the assumptions that can be reasonably made about the distribution of these data. This section is a brief guide to help reviewers identify correctly interpreted results for each of the methods discussed in this chapter.

(a) *Sign test.* This is used in the matched-pair situation. The conclusion that can be made is that there is a higher proportion of success in one condition (say, a treatment) than in the other condition (e.g., a control). If the researcher uses the more general binomial test, a confidence interval can be provided for the proportion (or percentage) of successes in the treatment condition when compared to the control. Any proportion over 0.5 (or percentage over 50%) is indicative of treatment success.

(b) *Goodness-of-fit test.* The goodness-of-fit test can lead to a conclusion that the response categories for the sampled population are present in proportions that differ from that of a norm group. If the researcher fails to reject this hypothesis, the conclusion is that there is not enough evidence to show a difference in populations. The researcher *cannot* state that the sampled population is the same as the norm group. In this case, the sample proportions form the best single guess about the proportional representation in the population.

(c) *One-sample Wilcoxon test.* If the data are quantitative and symmetrically distributed, the researcher can validly provide a confidence interval for the median. For matched pairs, symmetry of difference scores results when the two distributions of interest only differ in terms of location, and not shape. For example, this is the case if a treatment effect is consistent across the range of the distribution. In this situation, it is also valid for the researcher to state the confidence interval for

the mean instead of the median since the mean and median are the same when the distribution is symmetrical. If the research relies on ordinal data, or if the case for symmetry of difference scores is not compelling, then the researcher cannot provide a confidence interval, but a rejection of a null hypothesis due to higher treatment values can lead to the conclusion that the treatment condition is more likely to result in higher values than the control condition.

(d) *Irwin–Fisher test.* With samples that are too small to utilize the chi-square approximation, the conclusion is whether there is sufficient evidence to deem the proportion of successes higher in one condition than the other. If the chi-square approximation can be used, the researcher can additionally provide a confidence interval for the difference in the proportion success in the two conditions.

(e) *Mann–Whitney test.* If the data are quantitative and the two populations of interest have the same shape, such as when a treatment has a uniform effect across the range of values, then the researcher should provide a confidence interval for the median difference. The researcher should *not* refer to this as the difference in medians, but rather the median of differences. If the data are ordinal or if it is not reasonable to consider that the populations have the same shape, then with a rejection of the null hypothesis the researcher can conclude that higher values are more likely to come from the treatment condition than the control condition.

(f) *Kruskal–Wallis test.* This is an omnibus test, so rejection of the null hypothesis of median equality only leads to the conclusion that some populations are located higher on the scale than others. The researcher should conduct multiple comparisons to make more specific determinations about these differences. These comparisons lead to the same type of conclusions as the Mann–Whitney test under the same conditions described above. With quantitative data and similarly shaped distributions, confidence intervals can be provided for the median difference. Otherwise, the conclusions focus more generally on which distribution is located higher on the response scale.

(g) *Friedman test.* The conclusions for this within-subjects design parallel those of the Kruskal–Wallis test. Omnibus results simply point to differences in the conditions. Follow-up matched pair tests can yield confidence intervals for the median difference scores or statements about relative distribution location, depending on whether the data are quantitative and whether the researcher can make a reasonable case for the symmetry of difference scores, as described above in the discussion of the one-sample Wilcoxon test.

(h) *McNemar’s test.* If sample sizes are small, the researcher may only be able to claim a difference in the proportional outcomes of the two observational periods. With larger sample sizes, the researcher can add a confidence interval for the size of this difference.

(i) *Spearman’s correlation and Kendall’s tau.* For small sample sizes, a researcher using Spearman’s correlation coefficient can use a statistically significant result to conclude that there is a monotonic relation. Similarly, Kendall’s tau can be used to conclude an association of the variables based on a different ordering of judgments regarding two measures. When sample sizes are large enough, researchers can provide confidence intervals for the size of the association.

(j) *Tests of homogeneity and association.* These are omnibus tests. If the test statistic is sufficiently large, the researcher can conclude either that the populations have different distributions among the response categories or that there is an association between the two categorical response variables, depending on whether the test is for homogeneity or association. In either case, multiple comparisons should lead to confidence intervals for proportional differences.

13. Nonparametric Extensions Considered for More Complex Designs

This chapter refers to the most commonly used nonparametric methods to assist reviewers with the situations they are most likely to encounter. Yet the field of nonparametric statistics is broad, touching every possible research scenario. It extends to regression models, multivariate analysis,

structural equation models, multilevel models, and so on. Obviously these settings can lead to more complex analysis, so the best advice to a reviewer in these situations is to seek help from a statistician who specializes in nonparametric methods.

The Kruskal–Wallis and Friedman methods are two examples of rank-based methods for dealing with one-factor K -group designs, between-subjects and within-subjects, respectively. More complex designs with data that do not meet the criteria for traditional analysis of variance can be analyzed using aligned ranks procedures (Mehra & Sen, 1969). In brief, these methods “align” scores by removing effects to test additional effects prior to ranking the data. For example, to test the interaction of two factors in a two-factor completely randomized design, the researcher will subtract out the estimated effect of *Factor A* and *Factor B* from each of the observations, rank the resultant aligned scores, and then calculate a test statistic for the interaction effect. This general method enables the testing of effects within a broad range of models without requiring that the model residuals be normally distributed.

Finally, reviewers should be aware that although rank-based transformations are the most common in nonparametric analysis, other transformations can be used. For example, for any of the rank-based procedures discussed above, instead of replacing original scores with ranks, a researcher could replace the scores with normal scores. Normal scores are drawn from a normal distribution so that even if the original scores are not normally distributed, the transformed scores are normally distributed. This results in nonparametric methods that are at least as powerful as their parametric counterparts in every situation. That is, the power of the nonparametric test never drops below that of the corresponding parametric test. Aside from using a different transformation, the analytic methods are the same, regardless of whether the researcher has used ranks or normal scores, so the advice and cautions provided above remain the same in both situations.

References

- Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.
- Blair, R. C., & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various nonnormal distributions. *Journal of Educational Statistics*, 5, 309–335.
- Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin*, 97, 119–128.
- Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York: Wiley.
- Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, 20, 115–147.
- Hodges, J. L., Jr., & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t test. *Annals of Mathematical Statistics*, 27, 324–335.
- Hodges, J. L., Jr., & Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *Annals of Mathematical Statistics*, 33, 482–497.
- MacDonald, P. (1999). Power, Type I, and Type III error rates of parametric and nonparametric statistical tests. *Journal of Experimental Education*, 67, 367–379.
- Mehra, K. L., & Sen, P. K. (1969). On a class of conditionally distribution-free tests for interaction in factorial experiments. *Annals of Mathematical Statistics*, 40, 658–664.
- Meyer, J. P., & Seaman, M. A. (2014). A comparison of the exact Kruskal–Wallis distribution to asymptotic approximations for all sample sizes up to 105. *Journal of Experimental Education*, 81, 139–156.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156–166.
- Nussbaum, E. M. (2015). *Categorical and nonparametric data analysis*. New York: Routledge.
- Randles, R. H. (1980). Nonparametric statistical tests of hypotheses. In E. V. Hogg (Ed.), *Modern statistics: Methods and applications* (pp. 31–40). Providence, RI: American Mathematical Society.
- Sprent, P., & Smeeton, N. C. (2007). *Applied nonparametric statistical methods* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Srisukho, D., & Marascuilo, L. A. (1974). Monte Carlo study of the power of H-test compared to F-test when population distributions are different in form. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. Retrieved from www.eric.ed.gov.

27

Power Analysis

Kevin R. Murphy

One of the most common applications of statistics in the social and behavioral science is in testing null hypotheses. For example, a researcher wanting to compare two treatments will usually do so by testing the hypothesis that in the population there is no difference in the outcomes of these treatments. If this *null hypothesis* (H_0) can be rejected, the researcher is likely to conclude that there is a real (i.e., non-zero) difference in treatment outcomes. The power of a statistical test is defined as the probability that a researcher will be able to reject a specific null hypothesis when it is in fact false. One of the key determinants of power is the degree to which the null hypothesis is false; if treatments have a very small effect, for example, it may be difficult to reject the hypothesis that they have no effect whatsoever. Effect size is not the only determinant of power, however. The power of a statistical test is a complex nonlinear function of the sensitivity of the test, the nature of the treatment effect, and the decision rules used to define statistical significance.

There are several statistical models that have been used in defining and estimating the power of statistical effects. Kraemer and Thiemann (1987) derived a general model for statistical power analysis based on the intraclass correlation coefficient, and developed methods for evaluating the power of a wide range of test statistics using a single general table based on the intraclass correlation. Lipsey (1990) used the *t*-test as a basis for estimating the statistical power of several statistical tests. Murphy and Myors (1999) developed a model based on the noncentral *F* distribution and showed how it could be used with virtually all applications of the general linear model. Kelley and Rausch (2006) discussed methods of determining sample sizes needed to attain particular levels of accuracy in estimating a wide range of statistical parameters.

Cohen (1988), Lipsey (1990), and Kraemer and Thiemann (1987) provided excellent overviews of the methods, assumptions, and applications of power analysis. Murphy, Myors, and Wolach (2014) extended traditional methods of power analysis to tests of hypotheses about the size of treatment effects, not merely tests of whether or not such treatment effects exist. All of these sources describe the two main applications of power analysis, (1) in designing studies yet to be conducted (e.g., determining sample sizes, setting criteria for significance) and (2) in evaluating research that has already been completed (e.g., understanding why particular studies rejected or failed to reject the null hypothesis).

Desiderata for studies that apply power analysis methods are described in Table 27.1, and are explained in the sections that follow.

Table 27.1 Desiderata for Power Analysis.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. The hypotheses being tested are defined, alternative hypotheses are laid out, and analytic methods are chosen.	I, M
2. Three factors that determine statistical power—effect size, sensitivity, and decision criteria—are examined.	I, M
3. Statistical power is estimated, sample size requirements are determined, and decision criteria are evaluated.	M
4. The results of power analysis are reported.	M, R
5. Statistical power is considered in evaluating existing research and in planning future studies.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

1. Defining the Null Hypothesis

The first step in power analysis is to define the specific hypothesis that is being tested (the null hypothesis) and to define the *alternative hypothesis* (H_1) that will be accepted if a statistical test leads the researcher to reject the null. The term *null hypothesis* is typically used to refer to a specific point hypothesis (e.g., that the population difference between experimental and control conditions is zero) that can be tested and potentially rejected on the basis of data collected in a sample. There is considerable debate in behavioral and social sciences about the value and relevance of null hypothesis testing (Cohen, 1994; Cortina & Dunlap, 1997; Hagen, 1997; Harlow, Mulaik, & Steiger, 1997; Morrison & Henkel, 1970); one of the principal objections to this type of testing is that the null hypothesis that is most likely to be tested (e.g., that the population correlation between two variables is zero or that the difference between two treatments is zero) is often one that is very unlikely to be true (Murphy, 1990).

Numerous alternatives to traditional null hypothesis testing have been suggested. Serlin and Lapsley (1985) laid out procedures for creating and testing hypotheses that the effects of treatments or interventions are sufficiently close to those predicted on the basis of substantive theory to justify the conclusion that the theory is supported. Rouanet (1996) described Bayesian models for hypothesis testing. Murphy et al. (2014) described methods for forming and testing hypotheses that the effects of treatments or interventions exceed some minimum value. They also examined in detail the power of tests of the hypothesis that the effects of treatments are either trivially small or are large enough to be of substantive interest.

Statistical analysis should not normally be limited to tests of the traditional null hypothesis, in part because of the very low likelihood that this hypothesis is correct (Meehl, 1978; Murphy et al., 2014). At a minimum, studies that test traditional null hypotheses should also report information about the importance and accuracy of results (e.g., effect size estimates, confidence intervals). Alternatives to traditional null hypothesis tests (e.g., Bayesian tests, minimum-effect tests) should be carefully considered and used where applicable.

2. Factors that Affect Power

The power of a statistical test (e.g., a comparison of two sample means) is a function of its sensitivity, the size of the effect in the population, and the standards or criteria used to test statistical hypotheses. Tests have higher levels of statistical power when studies are designed to yield high levels of

sensitivity, when effect sizes (ES) are large, and/or when the criteria used to define statistical significance are relatively lenient. Studies should, if possible, be designed so that they achieve power levels of .80 or greater (i.e., so that they have at least an 80% chance of rejecting a false null hypothesis; Cohen, 1988; Murphy et al., 2014). When power is less than .50, it is not always clear whether statistical tests should be conducted at all, because of the substantial probability that they will fail to reject the null hypothesis even though it is false (i.e., commit a Type II error; rejecting the null hypothesis when it is true is referred to as a Type I error; Murphy et al., 2014).

Sensitivity refers to the ability of a study to consistently detect relatively small deviations from the null hypothesis. Researchers can increase sensitivity by using better measures (thereby reducing unsystematic variability associated with measurement error) or by using study designs that allow them to control for unwanted sources of variability in their data. The simplest and most common method of increasing the sensitivity of a study is to increase its sample size (n). Large samples should be used wherever possible; as n increases, statistical estimates become more precise and the power of statistical tests increases.

The formula for the standard error of the sample mean helps to illustrate how and why sample size affects sensitivity. The standard of the mean ($SE_{\bar{x}}$) is given by:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (1)$$

where σ is the population standard deviation of scores on the outcome measure. As this formula shows, as sample size (n) increases, the size of the standard error decreases, indicating smaller and smaller differences between population values and sample values (i.e., more precision). This formula also illustrates one of the most challenging barriers to achieving high levels of precision and sensitivity in a study, the curvilinear relation between sample size and precision. For example, doubling the sample size does not usually double the precision of sample estimates. Because the size of the standard error is usually a function of the square root of n , one implication is that in order to double the precision of estimates, one must often increase n by a factor of four. The nature of the relation between n and precision reinforces the recommendation that the largest possible samples be used.

The second factor that influences power is the size of the effect of treatments or interventions. If treatments have large effects, it is relatively easy to correctly reject the null hypothesis that the effect of treatments is zero, whereas small treatment effects might be difficult to detect reliably. Effect sizes are often measured in terms of statistics such as the standardized mean difference (d , the difference between treatment means divided by the pooled standard deviation) or the percentage of variance explained by treatments, interventions, and so forth (see Chapter 6 on effect sizes, this volume). Lipsey and Wilson (1993) reviewed effect sizes typically reported in the social and behavioral sciences. Effect size estimates might be obtained from meta-analyses of research literature (see Chapter 19,

Table 27.2 Some Conventions for Defining Effect Sizes.

	PV	r	d	f^2	Probability of a higher score in treatment group
Small effects	.01	.10	.20	.02	.56
Medium effects	.10	.30	.50	.15	.64
Large effects	.25	.50	.80	.35	.71

Sources: Cohen (1988), Grissom (1994)

Note: PV = percentage of variance explained; Cohen's $f^2 = R^2/(1-R^2) = \eta^2/(1-\eta^2) = PV/(1-PV)$, where $\eta^2 = SS_{\text{treatments}}/SS_{\text{total}}$

this volume), or might be derived from a substantive theory about the phenomenon being studied, but most power analyses depend on conventions of particular research communities to define the size of the effect that is expected or that is used in designing studies.

On the basis of surveys of research literature, Cohen (1988) suggested a number of conventions for describing treatment effects as “small,” “medium,” or “large” (see Table 27.2).

For example, a small effect might be described as one that accounts for about 1% of the variance in outcomes, or one where the treatment mean is about one-fifth of a standard deviation higher in the treatment group than in the control group. Similarly, a small effect can be described as one where the probability that a randomly selected member of the treatment group will have a higher score than a randomly selected member of the control group is about .56. In the absence of an acceptable estimate of the effect size expected in a particular study, it is common practice to assume that the effect will be small (e.g., $d = .20$; 1% of the variance accounted for by treatments) and to plan studies accordingly. Unless one has a good reason to believe that the effects of treatments are moderately large or larger, or only finds practical value in detecting effects of such magnitude, it is usually best to design studies so that they have sufficient power for detecting small effects. Studies designed to detect small effects will also have sufficient power for detecting larger effects.

Very large samples are often needed to yield adequate power for detecting small effects. For example, if the expected difference between control and treatment means is about one-fifth of a standard deviation (i.e., $d = .20$), a sample of 777 subjects will be needed to achieve power of .80 for rejecting the null hypothesis. In a study where subjects are randomly assigned to one of five treatments and where treatment differences are expected to account for 1% of the variability in outcomes, a sample of 1170 subjects will be needed to achieve this same level of power for an omnibus test of the hypothesis that all group means are identical. The smaller the expected effect of treatments or interventions, the more important it is to consider power in the design of studies.

Third, the power of statistical tests is affected by the standards or criteria used to define statistical significance, usually defined in terms of the alpha (α) level. Alpha is the conditional probability that a statistical procedure will reject a null hypothesis, given that this null hypothesis is in fact true; conventional approaches to null hypothesis testing define statistical significance criteria in such a way that the maximum value of α will be small (typically .05 or less). In the behavioral and social sciences, differences in treatment outcomes are usually regarded as statistically significant if the results obtained in a sample are outside of the range of results that would have been obtained in 95% of all samples drawn from a population in which the null hypothesis is true (i.e., $\alpha = .05$ is the most common, albeit arbitrary, threshold for “statistical significance” in the behavioral and social sciences). Some researchers use more stringent criteria when defining statistical significance, for example demanding that the alpha level should be set at .01 or smaller before sample results are declared statistically significant. The use of stringent criteria (e.g., $\alpha = .01$ or lower) for defining statistical significance is not recommended. Unless there are good reasons to believe that the null hypothesis might be true (because the null hypothesis is a point hypothesis, this is rarely the case; Cohen, 1994; Meehl, 1978), use of stringent criteria for defining statistical significance will normally lead to reductions in power without providing substantial benefits.

3. Power, Sample Size, and Criteria for Significance

The power of a null hypothesis test is a function of n , ES , and α , and the equations that define this relation can be easily rearranged to solve for any of four quantities (i.e., power, n , ES , and α), given the other three. The two most common applications of statistical power analysis are in: (1) *post-hoc* analysis – determining the power of a study, given n , ES , and α , and (2) *a-priori* analysis – determining how many observations will be needed (i.e., n required), given a desired level of power, an ES

estimate, and an α value. Both of these analyses are extremely useful in planning research, and are usually so easy to do that they should be a routine part of designing a study.

Both of these applications of power analysis assume that a decision has been made about the significance criterion to be used (e.g., $\alpha = .05$) and that there is some basis for estimating the size of treatment effects, or the degree to which the null hypothesis is likely to be wrong. It is best to be conservative in estimating ES s; as noted earlier, if you have no credible *a priori* basis for making this estimate, it is typically best to base power analyses on the assumption that treatment effects will be small. In some studies, observed treatment effects are used to make an estimate of population treatment effects, but such *post hoc* power analyses are usually discouraged (Hoening & Heisey, 1971), in part because they tend to provide overestimates of power.

Once ES has been estimated and α has been chosen, it is easy to either determine the power a particular study will provide or to determine the sample sizes needed to reach specific levels of statistical power. The equations that define the power of various statistical tests are not complex, but the calculation of power is somewhat tedious, and it is common practice to use power tables or power analysis software to perform the necessary calculations. Given a particular sample size (n) and α level, the power of a statistical test is a nonlinear monotonic function of the ES that asymptotes at or near 1.0. That is, the probability that a researcher will reject any particular null hypothesis approaches unity as the gap increases between the null hypothesis (e.g., that treatments have no effect) and the reality that they might have large effects. Cohen (1988) provided among the most complete sets of power tables readily available, based on calculations of power for a variety of different statistical tests, whereas Murphy et al. (2014) provided a smaller set of tables that can readily be adapted to most of the statistical tests discussed by Cohen.

An example might clarify the two main applications of power analysis. Suppose a researcher is comparing two different methods of instruction. Table 27.3 displays the results of a power analysis in two different ways, first showing the level of power for statistical comparisons ($\alpha = .05$) of two groups given various values of n and ES (expressed in terms of the percentage of variance in the dependent variable explained by treatments), then showing the number of subjects that would be needed to yield power of .80 given different ES values.

If the percentage of variance (PV) in outcomes explained by treatments is relatively small (e.g., $PV = .02$), a relatively large sample ($n = 387$) will be needed to attain power of .80. This power analysis shows that when the sample size is 200 and $PV = .02$, the probability that the null hypothesis will be (correctly) rejected is .50, suggesting that with a sample this large null hypothesis tests will essentially be a coin flip. On the other hand, when the effect of treatments is large (e.g., $PV = .10$), samples with $n = 75$ will have an 80% chance of correctly rejecting the null hypothesis.

Table 27.3 can also be used to estimate the types of effects that could be detected with a fixed level of power, given n and α . For example, assume that 100 subjects are available for a study and that the researcher desires a power of .80 or greater for statistical tests that employ a .05 α level. Table 27.3 makes it clear that this level of power will only be achieved if the ES is a bit greater than $PV = .05$, but less than $PV = .10$ (a PV of approximately .066 is required to achieve power of .80).

Table 27.3 Two Ways of Displaying the Outcomes of a Power Analysis.

Power Levels			<i>n</i> Required for Power of .80		
<i>n</i>	$PV=.02$	$PV=.05$	$PV=.10$	PV	<i>n</i>
100	.27	.61	.92	.02	387
200	.50	.91	.999	.05	153
500	.90	.999	.999	.10	75

Finally, power analyses might be used to aid in making rational decisions about the criteria used to define statistical significance (Cascio & Zedeck, 1983; Nagel & Neff, 1977). For example, suppose researchers are comparing two treatments with 200 subjects assigned to each treatment. The researchers expect a relatively small treatment effect of, say, $PV = .02$. Using $\alpha = .05$, power would be 0.64. If $\alpha = .01$ is used, power drops to 0.37 (Cohen, 1988). The trade-off between Type I error protection and power suggests that a researcher must balance risk and consequences of a Type I error with risk and consequences of a Type II error.

In general, the choice of a more stringent α level (e.g., choosing .01 rather than .05) will lead to reductions in power. This choice might make sense if researchers are more concerned Type I errors than with Type II errors. Cascio and Zedeck (1983) presented equations for estimating the relative weight given to Type I vs. Type II errors in various research designs, which can help researchers evaluate these tradeoffs. They showed that the apparent relative seriousness (ARS) of these errors implied by a study design can be estimated using:

$$\text{ARS} = \frac{p(H_1)(1-\text{power})}{(1-p(H_1))\alpha} \quad (2)$$

where $p(H_1)$ = probability that H_0 is false.

For example, if the researcher believes that the probability that treatments have *some* effect is .7, and $\alpha = .05$ and the power is .80, the choice of the $\alpha = .05$ significance criterion implies that a mistaken rejection of the null hypothesis (i.e., a Type I error) is 9.33 times as serious [i.e., $(.7 \times 2)/(.3 \times .05) = 9.33$] as the failure to reject the null when it is wrong (i.e., a Type II error). In contrast, setting $\alpha = .10$ leads to a ratio of 4.66 [i.e., $(.7 \times 2)/(.3 \times .10) = 4.66$], or to the conclusion that Type I errors are treated as if they are 4.66 times as serious as a Type II error (see also Lipsey, 1990).

The first advantage of Equation (2) is that it makes explicit the values and preferences that are usually not well understood, either by researchers or by the consumers of social science research. In the scenario described above, choice of an α level of .05 makes sense only if the researcher thinks that Type I errors are over nine times as serious as Type II errors. If the researcher believes that Type I errors are only four or five times as serious as Type II errors, he or she should set the significance level at .10, not at .05.

4. Reporting Results

There is no standard format for reporting the outcomes of a power analysis but it is relatively simple, on the basis of the known determinants of power, to determine what information should be reported. Because power is usually a function of three variables (i.e., n , ES , and α), it is best to include information about all three in discussions of statistical power. For example, one might report:

On the basis of our review of the literature, we expected that the difference between two treatments would correspond to a medium-sized effect (i.e., $d = .50$). Our study included 120 subjects who were randomly assigned to treatment and control conditions. We used a two-tailed test ($\alpha = .05$) to compare group means. The power of this test for detecting medium effects (e.g., $d = .5$), under assumed conditions (normality, independence of observations, homogeneity of variance), is .77.

It is also important to report the method or the statistical software used to estimate power. Power analyses are included as part of several statistical analysis packages (e.g., SPSS provides Sample Power, a flexible and powerful program) and it is possible to use numerous websites to perform simple power analyses. Some power analysis textbooks (e.g., Murphy et al., 2014) include software

for performing these analyses. Two notable software packages designed for power analysis are both available for free:

- *G*Power* (Faul, Erdfelder, Lang, & Buchner, 2007; www.gpower.hhu.de/en.html) available for both Macintosh and Windows environments. It is simple, fast, and flexible.
- *pwr*, a package in R (<https://github.com/heliosdrm/pwr>). This package is easy to install and use and is compatible with virtually all platforms

5. Evaluating Existing Research and Designing Future Studies

Power analyses are extremely valuable for understanding the outcomes of significance tests in the published literature. For example, suppose a researcher reports a statistically significant correlation between pre-employment drug tests and subsequent job performance. If the study uses very large samples, a statistically significant finding might not be very meaningful. If $n = 5000$, the power for detecting a correlation as low as .04 exceeds .80. Similarly, a researcher who reports that there is no statistically significant correlation between mothers' health and the health of children, based on a sample of $n = 30$, might lead readers seriously astray. If $n = 30$, power is less than .80 for detecting correlations as large as .45, and it is possible that a sample this small could miss a substantial correlation between these two variables. Both of the examples illustrate a key point about null hypothesis testing. Describing the correlation between two variables as "statistically significant" does not necessarily mean that it is large or important, and describing this same correlation as "statistically nonsignificant" does not necessarily mean that it is small and unimportant. It is recommended that whenever interpreting the results of statistical significance tests, power should be considered. When tests of the null hypothesis are carried out with either very high levels of power or very low levels of power, the outcomes of these tests are virtually a foregone conclusion, and the power of these tests should be routinely considered when evaluating their results.

Power analyses can be useful in understanding the likelihood that the results in a particular study will replicate well in future studies. For example, if a power analysis indicates that power is quite low (e.g., .60), it is still possible that a study will reject the null hypothesis. However, if the estimates of effect size are reasonably accurate, one should not expect that replications of that study will consistently reject the null hypothesis. On the contrary, if effect size estimates are accurate and there are ten replications of a study, the best guess is that only six of these will reject the null hypothesis. The parameter α is often misinterpreted as an indication of the probability that test results will replicate. Assessments of power are recommended as a much more accurate barometer of whether future tests conducted in similarly-designed studies are likely to lead to the same outcomes.

Power analyses can be quite useful in settings where one wants to argue that the null hypothesis is at least close to being correct. For example, if one wanted to argue that a new type of training has no real effects, one way to make this argument is to design a study that has a high level of power. If a powerful study is conducted and one still fails to reject the null hypothesis, one might conclude that this null is at least reasonably close as an estimate of the true state of affairs. Whenever researchers want to use the failure to reject the null hypothesis as evidence that this hypothesis is at least approximately true, they should first demonstrate that their studies have sufficient power to reject the null when it is meaningfully wrong.

Finally, power analysis should be carefully considered when designing future studies, particularly when making choices about sample size. There are times when practical constraints make it impossible to obtain the large samples needed to reliably detect small but potentially important effects. If there are constraints on the maximum sample size that can be attained, power analysis can be used to determine the type of effect that can be detected, given a fixed level of power, or the level of power that

can be attained given a fixed effect size. For example, suppose a research team is interested in comparing the effects of two drugs and uses a two-tailed *t* test (e.g., $\alpha = .05$) to determine whether there are differences in the drug effects. They expect a relatively small difference in the effectiveness of the drugs (e.g., $d = .15$), and a power analysis shows that a sample of at least 1398 will be needed to reach power of .80 for tests of the null hypothesis that the drugs have identical effects. They can afford to sample only 800 subjects. With a sample this large, they will have power of .80 or above for detecting somewhat larger effects ($d = .20$), and will have to make a decision about whether it is realistic to expect effects this large. If they are confident that the effect will be approximately $d = .15$ and they are limited to a sample of 800 participants, power for testing the null hypothesis will be only .56, suggesting that they almost as likely to conclude that there is no detectable difference between the drugs than they are to conclude that there is a small, but potentially important difference in the effects of the drugs.

In sum, power analysis is extremely useful as a tool for planning and evaluating research studies. Studies in the behavioral and social sciences are often conducted with low levels of power (Cohen, 1988; Maxwell, Kelley, & Rausch, 2008). Researchers who pay careful attention to power analysis are less likely to make Type II errors or to misinterpret the outcomes of null hypothesis tests.

References

- Cascio, W. F., & Zedeck, S. (1983). Open a new window in rational research planning: Adjust alpha to maximize statistical power. *Personnel Psychology*, 36, 517–526.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997–1003.
- Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, 2, 161–173.
- Paul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.
- Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, 79, 314–316.
- Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, 52, 15–24.
- Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.) (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum.
- Hoening, J. M., & Heisey, D. M. (1971). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55, 19–24.
- Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, 11, 363–385.
- Kraemer, H. C., & Thiemann, S. (1987). *How many subjects?* Newbury Park, CA: Sage.
- Lipsey, M. W. (1990). *Design sensitivity*. Newbury Park, CA: Sage.
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist*, 48, 1181–1209.
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, 59, 537–563.
- Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of psychology. *Journal of Consulting and Clinical Psychology*, 46, 806–834.
- Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy: A reader*. Chicago, IL: Aldine.
- Murphy, K. R. (1990). If the null hypothesis is impossible, why test it? *American Psychologist*, 45, 403–404.
- Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology*, 84, 234–248.
- Murphy, K. R., Myors, B., & Wolach, A. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (4th ed.). New York: Taylor & Francis.
- Nagel, S. S., & Neff, M. (1977). Determining an optimal level of statistical significance. *Evaluation Studies Review Annual*, 2, 146–158.
- Rouanet, H. (1996). Bayesian methods for assessing the importance of effects. *Psychological Bulletin*, 119, 149–158.
- Serlin, R. A., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, 40, 73–83.

28

Propensity Scores and Matching Methods

Elizabeth A. Stuart

Many studies aim to estimate causal effects of risk factors, interventions, or programs, on outcomes of interest. While randomization is generally seen as the preferred design for estimating causal effects it is not always possible to randomize the “treatments” of interest, especially in the social sciences. Propensity scores are a useful tool that can help yield better estimates of causal effects in non-experimental studies by ensuring that the treatment and comparison groups are similar with respect to the observed covariates. The propensity score itself is defined as the probability of receiving the treatment given a set of observed covariates (Rosenbaum & Rubin, 1983a). It is used to equate (or “balance”) the covariates between the treatment and comparison groups through propensity score matching, weighting, or subclassification (Stuart, 2010). Outcomes can then be compared between the equated groups, with less risk for extrapolation from treatment to comparison group (and vice versa), thus yielding more reliable causal effect estimates (Ho, Imai, King, & Stuart, 2007). For overviews and current methods, see Hernan and Robins (2015), Imbens and Rubin (2015), Stuart (2010), and Rosenbaum (2009).

Table 28.1 Desiderata for Propensity Score and Matching Methods.

<i>Desideratum</i>	<i>Manuscript section(s)*</i>
1. A clear statement of the causal question of interest is provided, including a definition of treatment and comparison conditions and outcomes of interest.	I, M, D
2. Clear specification of the estimand of interest is made.	I, M, R, D
3. The sample is clearly defined and how individuals were selected into the sample is specified.	M
4. A clear statement of the potential confounders is provided, and if (and how) they are measured in the data available. Justification for variables is included, taking care to exclude variables that may be affected by the treatment of interest.	M
5. Mention and provide some justification of the assumption of no unmeasured confounders.	M, D
6. Specification of the propensity score model, including estimation procedure, covariates, and interactions included, is provided.	M

- | | |
|---|---------|
| 7. Specification of how the propensity scores were used in the analysis is provided (i.e., details of matching, details of subclassification, or details of weighting), including a rationale for the use of that approach. | M |
| 8. Whether any covariates were handled in a nonstandard way, such as through exact matching, is explained. | M |
| 9. The name and version of the statistical software package used to conduct the propensity score analysis and estimate treatment effects is presented. | M |
| 10. If relevant, discussion is provided of other propensity score approaches used (e.g., sensitivity analyses), and how the primary approach was selected. | M, R |
| 11. The treatment of missing data (on treatment status, covariates, and/or outcomes) is addressed. | M |
| 12. Specification of the outcome model is provided, including whether covariates were adjusted for in the outcome analyses (e.g., in a doubly robust approach). | M |
| 13. Sample size in the treatment and comparison groups, before and after propensity score adjustment, is presented. | R |
| 14. A display of the extent of propensity score overlap between treatment and comparison groups is provided. | R |
| 15. Covariate balance before and after the propensity scores are applied, with evidence that balance reached an acceptable level, are presented. | R |
| 16. Assessment of sensitivity to a potential unobserved confounder is described, including how the sensitivity analysis was conducted. | M, R, D |
-

* I = Introduction, M = Methods, R = Results, D = Discussion.

1. Defining the Causal Effect of Interest

Propensity scores are a key tool for estimating causal effects in non-experimental settings. The first step in estimating a causal effect is to be very clear about the quantity of interest: what is the precise causal effect that a researcher aims to estimate? A causal effect is inherently a comparison of potential outcomes: the potential outcome that a unit (e.g., an individual) would have if they receive the treatment condition, generally denoted $Y(1)$, and the potential outcome that same unit would have under the comparison condition, denoted $Y(0)$. The causal effect of interest is often the difference between these, $Y(1) - Y(0)$. For each unit we can observe either $Y(0)$ or $Y(1)$ (not both), known as the “fundamental problem of causal inference” (Holland, 1986). Because estimating individual-level causal effects is generally very difficult, we often aim instead to estimate an average causal effect across some population of interest.

To define the causal effect, researchers need to clearly specify: (1) the treatment condition or exposure of interest (e.g., a new behavior program for first graders), (2) the comparison condition of interest (e.g., the “usual” behavior management program used in schools), (3) the relevant “units” (e.g., classrooms or students), and (4) the outcome(s) of interest (e.g., suspension rates in second grade). Specifying the comparison condition is just as important as the treatment condition. For example, when studying the effects of “heavy” marijuana use in adolescence, researchers need to decide whether they are interested in the effects of heavy use as compared to no marijuana use, or perhaps heavy use as compared to light use (Stuart & Green, 2008); each would be a valid comparison, but answering different scientific questions.

Early in any manuscript estimating causal effects, authors must clearly state the causal effect of interest, and justify it scientifically, providing theory or preliminary evidence regarding its plausibility and interest. In particular, the units, treatment, comparison condition, and outcome(s) must also be clearly stated and described. This is generally done in the Introduction to set up the main

point of the paper, but should also be returned to in the Methods section (in particular to provide full details on the specification of key constructs, such as the treatment, covariates, and outcomes), and in the Discussion section, in particular to connect back to the broader research literature and consider ideas for future work.

2. Clear Specification of the Estimand of Interest

There are two additional considerations for describing the causal estimand of interest in a non-experimental study. The first is the specific comparison of potential outcomes that is of interest. With continuous covariates this is often the difference in potential outcomes, $Y(1) - Y(0)$. For binary covariates this could be an odds ratio, relative risk, or risk difference (Austin, 2010).

An additional important clarification, though, concerns the population of interest. This is particularly important to specify because different propensity score methods sometimes estimate different quantities. One common quantity of interest is the “average treatment effect” (ATE), defined as the average difference in potential outcomes, across the n individuals in the treatment *and* comparison

groups (the combined population): $ATE = \frac{1}{n} \sum_{i=1}^n [Y_i(1) - Y_i(0)]$. In some cases, the average treatment effect on the treated (ATT) is of interest; this is the same quantity as the ATE, but averaged just over individuals in the treatment group, where T_i is the treatment assignment indicator for the i th unit:

$$ATT = \frac{1}{\sum_{i=1}^n T_i} \sum_{i=1}^n T_i * [Y_i(1) - Y_i(0)].$$

Data availability and overlap between treatment and comparison groups may dictate which estimand it is possible to estimate (see, e.g., Crump, Hotz, Imbens, & Mitnik, 2009), but researchers should start out with an idea of the estimand they would like to estimate. The ATT is often particularly relevant for treatments or interventions believed to be potentially harmful, such as adolescent drug use: we would never think to impose drug use on a group of adolescents, but rather are potentially interested in learning about comparing outcomes between a group of adolescents who were heavy drug users with what would have happened to those same adolescents, had they not been heavy drug users. In contrast, the ATE may be more relevant for policies or interventions that may be delivered or received by an entire population of people. Any manuscript describing the results from a non-experimental study should clearly state throughout what the estimand of interest is. The method used should be selected to correspond to the estimand of interest, and the Results and Discussion sections should describe the results in a way consistent with the estimand of interest (see, e.g., DuGoff, Schuler, & Stuart, 2014).

3. Clear Definition of Sample

Any manuscript should clearly describe in the Methods section how individuals were selected to be in the study, and whether any exclusions were applied. This includes how the original sample was obtained as well as any resulting exclusions, such as for missing data or because of focused interest in a particular sub-sample (e.g., males, or youth between 13 and 18 years old). A full accounting of the sample should include something like a flow chart indicating the original sample and then the reasons individuals were excluded, leading to the final analysis sample.

The sample needs to be defined in a way that does not depend on “post-treatment” variables. For example, in a propensity score analysis inclusion in the sample should not be related to the value of their outcome of interest, or of any other mediators of interest. Sample inclusion should be a function only of things measured at baseline, before the treatment was applied. While this is easy to do in clean longitudinal settings (e.g., a longitudinal study with covariates collected before the treatment is applied, and outcomes measured after), many studies using propensity score methods use

cross-sectional data from a single point in time. In such studies researchers need to think carefully about temporal ordering and ensure the appropriate ordering as much as possible. For example, it would be inappropriate to control for third grade test scores in a study of the effects of preschool experiences on high school graduation rates.

4. Details on Observed Confounders

The propensity score itself is the predicted probability of receiving the treatment, given the observed covariates included in the propensity score model. Manuscripts describing propensity score analyses should therefore include a clear listing of the covariates included in the propensity score model. The Methods section should detail the variables themselves, how they are constructed, and how they were measured.

As with defining the sample of interest it is important to keep temporal ordering, so all covariates should be measured before the treatment or comparison condition is applied. Thus, description of the variables should include specification of when they were measured, or why they are likely not affected by the treatment. Including variables that were affected by the treatment may lead to “post-treatment bias” and incorrect conclusions regarding the causal effects (Frangakis & Rubin, 2002; Greenland, 2003). Researchers interested in controlling for post-treatment variables (such as level of compliance or a mediator) should use methods appropriate for those analyses, such as principal stratification (Frangakis & Rubin, 2002) or causal mediation analysis (Jo, 2008; Keele, 2015; VanderWeele, 2015).

5. Assumption of No Unmeasured Confounders

The key assumption underlying studies using propensity score methods is that of no unmeasured confounding, also known as ignorability, unconfoundedness, or “no hidden bias” (Rosenbaum, 2002; Rosenbaum & Rubin, 1983a). This assumption states that, given the observed covariates X (those included in the propensity score or matching process), treatment assignment (T) is independent of the potential outcomes:

$$T \perp (Y(0), Y(1)) | X$$

If this assumption holds, after adjusting for the observed covariates (through using the propensity score), the difference in outcomes between treatment and comparison groups can be interpreted as a causal effect. For example, if 1:1 propensity score matching was done and if ignorability holds, we can treat the matched samples as if they came from a randomized trial.

This is clearly an important assumption that may not always be satisfied. Any manuscript using propensity score methods should discuss this assumption in the Methods section (as the key underlying assumption of the method used), and provide discussion of the plausibility of this assumption in the Discussion section. The plausibility can be enhanced by including a large set of covariates in the propensity score approach (see, e.g., Shadish, Clark, & Steiner, 2008; Steiner, Cook, Shadish, & Clark, 2010). To assess the plausibility it is important for researchers to have a good understanding of the treatment assignment mechanism: what factors went into the choice or decision of which treatment individuals received, and then, ideally, to have all of the important factors observed (see Desideratum 16 for more details). This is a place where mixed methods studies can be of particular value, to combine qualitative information about the treatment assignment process with the quantitative propensity score approach.

6. The Propensity Score Model

The specification and estimation procedure used to estimate the propensity score itself should be provided in the Methods section of any manuscript using propensity scores. The estimation procedures can be parametric, such as logistic regression, or non-parametric, such as generalized boosted models (McCaffrey, Ridgeway, & Morral, 2004) or random forests; in fact, recent work has found that non-parametric procedures can work better for propensity score estimation (Lee, Lessler, & Stuart, 2010). The estimation procedure should be described. If a parametric model is used, the specification should be provided, including the covariates included and whether any interactions were included in the model. If a non-parametric model is used, the specification of that should be provided (e.g., the number of trees or other stopping rules), including a listing of the covariates used in the procedure.

7. How the Propensity Scores Were Used

There are many ways of using propensity scores; broadly, the methods fall into three categories: matching, weighting, and subclassification (Stuart, 2010). Matching typically selects for each treated individual a number (often one, but sometimes 2-3) of comparison individuals with similar propensity scores. Full matching or variable ratio matching uses a similar idea, but allows the number of comparison subjects matched to each treated subject to vary (Stuart & Green, 2008). Weighting constructs a weight for each unit that is a direct function of the propensity score. For example, the most common weighting approach, inverse probability of treatment weighting (IPTW; Lunceford & Davidian, 2004), gives a weight of one over the propensity score (the probability of being treated) to the treatment group members, and a weight of one over one minus the propensity score (the probability of being in the comparison group) to the comparison group members to estimate the average treatment effect (ATE). In this way both groups are weighted to the combined sample of treatment and comparison group members. (An alternative weighting strategy, sometimes called “weighting by the odds,” can be used to estimate the ATT; see Hirano, Imbens, & Ridder, 2003). Finally, subclassification proceeds by creating bins (subclasses) of individuals based on their propensity score values and estimates treatment effects separately within each subclass before averaging across subclasses. The quintiles or deciles of the propensity score distribution are often used to define the subclasses.

In any manuscript using propensity score methods it is crucial for authors to specify how the propensity scores were used. This includes describing the general approach (e.g., matching, weighting, or subclassification), as well as any details specific to the approach (e.g., if matching is used, was it matching with or without replacement, how many matches selected for each treated subject; for subclassification, details such as how the subclasses were defined (e.g., quintiles of the propensity score) need to be provided). If weighting is used, authors should specify whether any trimming was implemented (Lee, Lessler, & Stuart, 2011).

The approach used should also connect back to the estimand of interest. Matching methods typically estimate the average treatment effect on the treated. Subclassification and weighting can estimate either the average treatment effect or the average treatment effect on the treated, depending on exactly how those approaches are implemented (e.g., how the weights are created).

Authors should also discuss how units outside the range of common support were handled, for example if subjects were discarded if they did not have a good match or if, *a priori*, analysis was restricted to those individuals with propensity score values overlapping that of the other group, or if calipers were used in the matching procedure (see, e.g., Crump et al., 2009; Dehejia & Wahba, 1999; King & Zeng, 2006). Further, the authors should provide some rationale for the use of a particular approach, perhaps related to the estimand of interest, the performance of different methods in terms of the balance achieved, and/or the relative sample sizes of the treatment and comparison groups.

8. Handling of Specific Covariates

In some settings it makes sense to try to get particularly good balance on key covariates—perhaps those that will be investigated as effect moderators or those believed to be strongly related to outcomes (Rubin & Thomas, 2000). This can be accomplished by stratifying the data on that variable and conducting the propensity score approach separately within groups (e.g., by gender, as in Stuart & Green, 2008), or through matching procedures that are designed to get particularly good balance on a small set of covariates while still providing good balance on a larger set (e.g., Mahalanobis metric matching within propensity score calipers; Rubin & Thomas, 2000). If certain variables were handled in a special way that should be clearly stated in the text.

9. Software

The authors should clearly specify the statistical software package used to implement the propensity score approach used and to estimate the treatment effects, with enough detail for readers to be able to replicate the analysis.

10. Alternative Approaches Considered

In the Methods (and possibly Results) sections, authors should describe any other propensity score approaches attempted or used, as well as a justification for why the primary approach was selected. The choice often involves an assessment of the propensity score overlap between the treatment and comparison groups (also known as common support), and potentially the resulting effective sample size resulting from the approach.

A common way of selecting a particular propensity score approach is to use the approach that provides the best covariate balance (see, e.g., Harder, Stuart, & Anthony, 2010; Hill, Rubin, & Thomas, 1999). If authors used this strategy they should at least briefly mention the propensity score approaches tried, and the metric(s) used to select the primary approach. If multiple approaches yield good covariate balance it may make sense for authors to present results from multiple approaches as sensitivity analyses. To maintain the separation of “design” and “analysis” (Rubin, 2001, 2007), the choice of method should *not* depend on the outcome values or resulting treatment effect estimates.

11. Missing Data

As with nearly any analysis, authors using propensity score methods should clearly discuss missing data, including the extent of missingness on the treatment indicator, covariates, and outcome(s). There should also be a clear statement of how the missing values were handled, for example, through mean imputation, single or multiple imputation, or a missing data indicator approach (which can be appropriate for propensity score estimation; Rosenbaum & Rubin, 1984). Mitra and Reiter (2016) and Qu and Lipkovich (2009) discussed methods for integrating propensity score methods and multiple imputation.

12. Outcome Model Specification

In the Methods section, authors using propensity score methods also need to state clearly how the outcome analysis was conducted. Generally, for nearest neighbor matching methods, the outcome analysis is conducted in the matched samples. For weighting, outcome models are run with the propensity score weights. For subclassification, effects are estimated within each subclass and then

aggregated across subclasses (Lunceford & Davidian, 2004). If subclassification is used, authors should specify how effects were averaged across subclasses to obtain an overall effect estimate.

Authors also need to clarify in the Methods section whether the outcome models adjusted for any of the covariates. It is common to include the covariates in the propensity score model and the outcome model, in the spirit of “doubly robust” models (Bang & Robins, 2005; Ho et al., 2007; Rubin, 1973). If only a subset of the covariates are included in the outcome model (e.g., those with relatively poor balance after the propensity score approach) that should be clarified for the reader.

13. Sample Sizes

The sample sizes in the treatment and comparison groups, both before and after the propensity score approach is implemented, needs to be clearly presented in the Results section, and connected back to the original study sample. This presentation should include clarification regarding any cases dropped, either from the original study sample (e.g., individuals with missing treatment or outcome values), or dropped from the study sample as a result of the propensity score approach utilized (e.g., those outside the range of common support, or those without a match within their caliper).

14. Display of Propensity Score Overlap

Authors should present in the Results section (or an Appendix) a graphical display of the propensity score distributions in the treatment and comparison groups, and their overlap, ideally both before and after the propensity score approach is applied. This graphical display could be a histogram, density plot, or boxplot; any graphic that allows the comparison of distributions between the two groups, and the extent of overlap.

15. Covariate Balance

The author should provide in the Results section evidence that the propensity score approach “worked” in terms of creating groups with similar covariate distributions. Because *p*-values from Kolmogorov-Smirnov or *t*-tests can be misleading due to either changes in balance OR to changes in effective sample size (Imai, King, & Stuart, 2008), a more common metric for covariate balance is the standardized difference in means. Similar to an effect size, this metric is calculated by dividing the difference in means between the treatment and comparison groups by the standard deviation of the covariate, thus putting all of the differences on a standard deviation scale. This same formula can be used for binary covariates, or a simple difference in proportions be presented (Austin, 2009). These metrics can be presented in table format, showing the balance before and after the propensity scores are applied, or in a graphical format that allows the reader to see the balance on each covariate before and after (e.g., a “Love plot,” as illustrated in Rudolph, Stuart, Glass, & Merikangas, 2014).

Of course the smaller the standardized differences in means the better (as close to 0 as possible), and there is also a general agreement in the propensity score literature that researchers should aim for a standardized difference in means less than 0.1 or 0.2 (Stuart, 2010). This recommendation is based on simulation studies in Rubin (1973) and Cochran and Rubin (1973), and discussion in Rubin (2001), which show that differences larger than this lead to more extrapolation between treatment and comparison groups, and more reliance on the parametric model forms. Of course this rule of thumb needs to be assessed in any particular study, and, for example, there may be desire to have even better balance on covariates believed to be particularly strongly related to the primary study outcomes.

16. Sensitivity to an Unobserved Confounder

The key assumption underlying propensity score analyses is that there is no unobserved confounder once the groups have been equated with respect to the observed covariates; that is, that treatment assignment is “ignorable” given the observed covariates, and that there is no hidden bias (Rosenbaum & Rubin, 1983a). Methods of assessing sensitivity to this assumption have been developed and generally ask the question “how strongly related to treatment and outcome would some unobserved confounder have to be to change the study conclusions?” See Rosenbaum and Rubin (1983b), Rosenbaum (2002), VanderWeele and Arah (2011), or Liu, Kuramoto, and Stuart (2013) for more details on these methods. Researchers using propensity score approaches to estimate causal effects in non-experimental settings should describe in the Methods section whether they have conducted any of these sensitivity analyses (and how), and then present the results in the Results section, putting it into context in the Discussion section. In the methods section, the sensitivity analysis approach used should be described (e.g., Rosenbaum bounds), and any needed input parameter values specified (as well as where the values for those came from).

Acknowledgments

Support for Dr. Stuart’s time on this project came from the US Department of Education Institute of Education Sciences (R305D150001; PIs Stuart and Dong).

References

- Austin, P. C. (2009). Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in Statistics-Simulation and Computation*, 38, 1228–1234.
- Austin, P. C. (2010). The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine*, 29, 2137–2148.
- Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–973.
- Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 417–446.
- Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1), 187–199.
- Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94, 1053–1062.
- DuGoff, E. H., Schuler, M., & Stuart, E. A. (2014). Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Services Research*, 49, 284–303.
- Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58, 21–29.
- Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*, 14, 300–306.
- Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, 15, 234–249.
- Hernan, M., & Robins, J. (2015). *Causal inference*. Boca Raton, FL: Chapman & Hall/CRC.
- Hill, J., Rubin, D. B., & Thomas, N. (1999). The design of the New York School Choice Scholarship Program evaluation. In L. Bickman (Ed.), *Research designs: Inspired by the work of Donald Campbell* (pp. 155–180). Thousand Oaks, CA: Sage.
- Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71, 1161–1190.
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15, 199–236.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–960.
- Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A*, 171, 481–502.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. New York: Cambridge University Press.
- Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods*, 13, 314–336.
- Keele, L. (2015). Causal mediation analysis: Warning! Assumptions ahead. *American Journal of Evaluation*, 36, 500–513.
- King, G., & Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, 14, 131–159.

- Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29, 337–346.
- Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PLoS ONE*, 6, e18174.
- Liu, W., Kuramoto, S. J., & Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in non-experimental prevention research. *Prevention Science*, 14, 570–580.
- Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23, 2937–2960.
- McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9, 403–425.
- Mitra, R., & Reiter, J. P. (2016). A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical methods in medical research*, 25(1), 188–204.
- Qu, Y., & Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine*, 28, 1402–1414.
- Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer.
- Rosenbaum, P. R. (2009). *Design of observational studies*. New York: Springer Verlag.
- Rosenbaum, P. R., & Rubin, D. B. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Rosenbaum, P. R., & Rubin, D. B. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(2), 212–218.
- Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516–524.
- Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29(1), 185–203.
- Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2, 169–188.
- Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26, 20–36.
- Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95, 573–585.
- Rudolph, K. E., Stuart, E. A., Glass, T. A., & Merikangas, K. R. (2014). Neighborhood disadvantage in context: The influence of urbanicity on the association between neighborhood disadvantage and adolescent emotional disorders. *Social Psychiatry and Psychiatric Epidemiology*, 49, 467–475.
- Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103, 1334–1344.
- Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15, 250–267.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25, 1–21.
- Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, 44, 395–406.
- VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. New York: Oxford University Press.
- VanderWeele, T. J., & Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology (Cambridge, Mass.)*, 22, 42–52.

29

Reliability and Validity

Ralph O. Mueller and Thomas R. Knapp

Both reliability and validity are essential parts of the psychometric properties of a measuring instrument.¹ The reliability of an instrument is concerned with the consistency of measurements: from time to time, from form to form, from item to item, or from one rater to another. On the other hand, the validity of an instrument is usually defined as the extent to which the instrument actually measures “what it is designed to measure” or “what it purports to measure,” that is, it assesses the relevance of an instrument for addressing a study’s purpose(s) and research question(s). Both reliability and validity are context-specific characteristics: for example, researchers are often interested in gauging if a measure remains reliable and valid for a specific culture, situation, or circumstance (e.g., a psychological test might be highly reliable and valid in a population of Caucasian adults but not in one of African American children). The conceptualization and specific definitions of reliability and validity have changed over time, as reflected in the various editions of Educational Measurement (Cronbach, 1971; Cureton, 1951; Feldt & Brennan, 1989; Haertel, 2006; Kane, 2006; Messick, 1989; Stanley, 1971; Thorndike, 1951). Such changes have also been reflected in the most recent edition of Standards for Educational and Psychological Testing (American Educational

Table 29.1 Desiderata for Reliability and Validity of Instruments.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. Each instrument used in the study is described in sufficient detail. The appropriateness of the instrument to address the study’s purpose(s) and research question(s) is made explicit.	I, M
2. Appropriate reliability indices are considered: The study’s purpose(s) guide the choice of indices calculated from current data and/or examined from previous research.	M, R
3. Suitable validity evidence is gathered: The study’s purpose(s) determine the type of validity support gathered from current data and/or consulted from related literature.	M, R
4. Applicable reliability and validity evidence is reported and interpreted. The study’s conclusions are placed within the context of such evidence (or lack thereof).	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Table 29.1 contains a list of desiderata regarding reliability and validity of instruments that should be followed in any empirical research report.

1. Instrument Description and Justification

Empirical data for analysis during research studies are collected with the aid of measuring instruments, be they laboratory equipment or, more common in the social and behavioral sciences, surveys, achievement batteries, or psychological tests. Because study results should only be trusted when investigators collect good data, authors should ensure that readers can judge the “goodness” of the data for themselves. Thus, at a minimum, a full description of the instrument(s) used is necessary (and should be followed by an assessment of reliability and validity; see Desiderata 2 and 3, respectively), including the purpose(s) and intended use(s), item format(s), and scales of measurement (i.e., nominal, ordinal, interval, or ratio). Obviously, a specific instrument is appropriate for use in some contexts but not necessarily in others (e.g., a high school reading test is likely to be inappropriate to measure a middle schooler’s intelligence). Authors must take care in justifying the choice of instrument(s) and making explicit the link to the study’s purpose(s) and research question(s). Often, the description and justification for a particular instrument is presented in a manuscript’s Instrumentation subsection of the Methods section but could also be accomplished in the Introduction.

2. Reliability Indices

Several approaches exist to assess an instrument’s reliability, dependent on a study’s specific purpose(s). Four traditional strategies often found in the literature are briefly discussed below: *test-retest*, *parallel forms*, *internal consistency*, and *rater-to-rater*. All four are based on classical test theory, but alternative conceptualizations of reliability exist that are based on other analytical frameworks: generalizability theory (see Chapter 9, this volume), item response theory (see Chapter 11, this volume), and structural equation modeling (see Chapter 33, this volume).

Test-retest reliability. If a study’s purpose is to assess measurement consistency of one instrument from one time point to another, a straightforward way to collect reliability evidence is to measure and then re-measure individuals and determine how closely the two sets of measurements are related (i.e., the test–retest method). In studies assessing psychological constructs such as attitude, a question with often serious ramifications is how much time should be allowed between the first and second testing. If the interval is too short, measurement consistency might only be due to the fact that individuals being tested “parrot back” the same responses at Time 2 that they gave at Time 1. If the interval is too long, some items might no longer be developmentally appropriate (e.g., academic achievement items on a middle school test administered to students entering high school) which could impact the validity of the instrument as well. Even in studies with physical variables, the length of time between measurements might be crucial for some (e.g., repeated weight measurements during a health awareness program might fluctuate widely, depending on weight loss/gain) but not for other variables (e.g., repeated height measurements of adult participants are likely to remain consistent, irrespective of the time-lag between measurements). In general, authors should justify their choice of time intervals between measurements as the acceptable amount of time is situation specific and somewhat subjective. An assessment of test–retest reliability can be accomplished in either an absolute manner (e.g., the median difference between corresponding measurements) or a relative manner (e.g., the correlation between the two sets of measurements), with the latter approach being more common than the former.

Parallel forms reliability. If a measuring instrument is available in two parallel (i.e., psychometrically equivalent and interchangeable) forms, say Form A and Form B, with measurements having been taken on both forms, reliability evidence can be obtained by comparing the scores on Form A with the scores on Form B, again either absolutely or relatively. The time between the administrations of the two forms is still an important consideration, but because the forms are not identical there is no longer the concern for “parroting back” if the time interval is short.

Internal consistency reliability. Given the disadvantage of multiple test administrations for test-retest and parallel forms reliability (e.g., increased costs, time lag between measurements, and missing data due to non-participation in second testing), a commonly used alternative is the estimation of the internal consistency of an instrument. Here, an instrument consisting of multiple items measuring the same construct is administered only once, but now treating the items as forming two parallel halves of the instrument. The two half-forms are created after the actual measurement, traditionally by considering the odd-numbered items as one form and the even-numbered items as the other form (though other ways to split an instrument are certainly possible, e.g., random assignments of items to halves). The scores on the two forms are then compared, usually relatively, by computing the correlation between the scores on the odd-numbered items with the scores on the even numbered items. This correlation must then be adjusted by using the Spearman-Brown formula (Brown, 1910; Spearman, 1910) in order to estimate what the correlation would have been between two full-forms, as opposed to two half-forms. That estimate is obtained by multiplying the correlation coefficient by two and then dividing that product by one plus the correlation. The type of reliability evidence thus produced is strictly concerned with internal consistency (from half-form to half-form) since time has not passed between obtaining the first set of measurements and obtaining the second set of measurements.

Another type of internal consistency reliability is from item to item within one form. Such an approach was first advocated by Kuder and Richardson (1937) for dichotomously scored test items and was subsequently extended to the more general interval measurement case by Cronbach (1951). Their formulae involve only the number of items, the mean and variance of each, and the covariances between all of the possible pairs of items. Cronbach called his reliability coefficient alpha. It is still known by that name and is by far the most commonly employed indicator of the reliability of a measuring instrument in the social sciences.²

Rater-to-rater reliability. When the data for a study take the form of ratings from scales, the type of reliability evidence that must be obtained before such a study is undertaken is an indication of the extent to which a rater agrees with him- or herself (intra-rater reliability) and/or the extent to which one rater agrees with another (inter-rater reliability). Several options exist to assess rater-to-rater consistency, with the intraclass coefficient and Cohen’s kappa (1960) being among the most popular (see Chapter 10, this volume).

Norm-versus criterion-referenced settings. Literature devoted to reliability assessment within norm-referenced versus criterion-referenced frameworks is plentiful. Most users of norm-referenced tests—where scores are primarily interpreted in relation to those from an appropriate norm or comparison group—have adopted approaches to reliability assessment similar to those summarized thus far, with particular emphasis on correlations that are indicative of relative agreement between variables. Criterion-referenced (or domain-referenced) measurement is concerned with what proportion of a domain of items has been answered successfully and whether or not that portion constitutes a “passing” performance (e.g., “John spelled 82% of the words on a spelling test correctly, which was below the cut point for progressing to the next lesson.”). Here, reliability assessment concentrates on measurement errors in the vicinity of the cut point with a particular interest in the reliability of the pass-or-fail decision. In parallel-form situations, for example, the matter of whether a person passes both Form A and Form B or fails both Form A and Form B, takes precedence over how high the correlation is between the two forms.

3. Validity Evidence

In physical science research the usual evidence for the validity of instruments is expert judgment and/or validity-by-definition with respect to a manufacturer's specifications, for example, "For the purpose of this study, body temperature is the number of degrees Fahrenheit that the Smith thermometer reads when inserted in the mouths of the persons on whom the measurements are being taken." As evidence of validity, the researcher might go on to explain that the Smith thermometer is regarded as the gold standard of temperature measurement.

In the social and behavioral sciences, investigators are often urged to provide evidence for *content validity* (expert judgments of the representativeness of items with respect to the skills, knowledge, etc. domain to be covered), *criterion-related validity* (degree of agreement with a "gold standard"), and/or *construct validity* (degree of agreement with theoretical expectations) of the measuring instruments used in their substantive studies. More recently, all three validity types have been subsumed under an expanded concept of *construct validity*, but not without controversy. Whatever conceptualization is used, researchers must be clear that instrument validity is not context free: a measure might be valid in one situation or for one population but not in or for another (e.g., the Scholastic Aptitude Test, SAT, is often argued as being valid to assess high school seniors' potential for success in undergraduate higher education but not for measuring their intelligence or potential to succeed in vocational training).

Criterion-related validity. When a measure is designed to relate to an external criterion, its validity is judged by either *concurrent* or *predictive* assessments (i.e., degrees to which test scores estimate a specified present or future performance). For example, a passing score on a driver's permit test with acceptable concurrent validity will allow the test taker to immediately drive motor vehicles, assuming an associated road test has been passed. On the other hand, evidence of predictive criterion-related validity is often helpful for judging instruments that are designed to measure aptitude with passing achievement scores serving as the standards for whether or not the aptitude tests are predictive of achievement. But, herein also lies an interesting dilemma: How does one know that the achievement tests themselves are valid? Do the standards need to be validated against an even higher standard? Or, if the standards' validity is established by expert judgment, why not appeal to experts directly for validity assessments of the aptitude measure? Furthermore, if expert judgment is to be the ultimate arbiter, who are the experts and who selects them?

Construct validity. In order to judge the degree to which a theoretical construct accounts for test performance, a researcher must assess the test's construct validity. Supportive evidence usually comes from exploratory or confirmatory factor analyses (see Chapter 8, this volume) in which the dimensionality and the degree of correlation of the variables comprising the instruments are investigated. The most popular approach is the *convergent/discriminant* strategy first recommended by Campbell and Fiske (1959): researchers determine the extent to which measurements obtained with the instruments in question correlate with variables with which they are theoretically expected to correlate (convergent) and the extent to which those measurements correlate with other variables with which they are theoretically not expected to correlate (discriminant). See also Chapter 24, this volume, on multitrait–multimethod analysis.

4. Reporting and Interpreting Reliability and Validity Results

Before reporting a study's main findings, investigators should discuss evidence of the reliability and validity of the instrument(s) used. Ideally, such evidence should come from both a thorough search of the related literature and an assessment based on current study participants. A comparison of present reliability and validity information with that gleaned from related literature is helpful to readers, especially when such information might be contradictory, as, for example, when earlier reliability/validity evidence could not be reproduced based on the current sample's data.

Certainly, when no previous reliability and validity information is available—as is the case when investigators construct their own instruments—authors must report psychometric properties of the instrument(s) based on an analysis of the current data. But even if reliability/validity evidence is identified from previous studies, it is often the case that it does not generalize to the current population under study. Thus, it is incumbent upon each investigator to provide a thorough justification for why the instruments used are appropriate for the current sample of participants.

In the social and behavioral sciences, reliability and validity coefficients in the .70 or .80 or above range are often considered acceptable with values below these cut-offs being acknowledged as study limitations. However, the acceptability of coefficients should be judged with caution as value adequacy depends on the particular phenomenon under study and whether the focus is on individuals or groups (see Reliability Standards 2.18 and 2.19 in American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Nevertheless, the interpretation of main results should commence from within the context of reliability and validity results as unreliability and/or invalidity usually attenuate the magnitude of expected findings and lead to wider confidence intervals and less likelihood of the detection of effects and relations in the data.

Notes

- 1 This chapter does not address the internal and external validity and reliability of a chosen research design (but see Chapter 30, this volume). Also, in fields outside the social and behavioral sciences, validity and reliability are sometimes known by different names. For example, in epidemiology, *reproducibility* is generally preferred over the term reliability. In engineering and related disciplines, equipment is said to be reliable if it does not, or is very unlikely to, break down. Also, the ambiguous term *accuracy* is sometimes used in lieu of either reliability or validity.
- 2 Kuder and Richardson actually derived several formulae for internal consistency by making successively relaxed assumptions and numbered them accordingly. The formula that is most frequently used to compute Cronbach's alpha is actually a direct extension of Kuder and Richardson's Formula Number 20 for dichotomous data.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi-trait multi-method matrix. *Psychological Bulletin*, 56, 81–105.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.
- Cureton, E. F. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education/Praeger.
- Kane, M. L. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151–160.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 171–195.
- Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 356–442). Washington, DC: American Council on Education.
- Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560–620). Washington, DC: American Council on Education.

30

Research Design

Sharon Anderson Dannels

The definition of research design is deceptively simple: it is a plan that provides the underlying structure to integrate all elements of a quantitative study so that the results are credible, free from bias, and maximally generalizable. “Research design provides the glue that holds the research project together” (Trochim, 2006, Design, ¶1). The research design determines how the participants are selected, what variables are included and how they are manipulated, how data are collected and analyzed, and how extraneous variability is controlled so that the overall research problem can be addressed. Regardless of the sophistication of the statistical analysis, the researcher’s conclusions may be worthless if an inappropriate research design has been used. Thus, design decisions both constrain and support the ultimate conclusions (Miles & Huberman, 1994).

Research designs may be identified as a specific design (e.g., a pretest-posttest control group design or a nonequivalent control group design) or by the broader category of experimental, quasi-experimental, or nonexperimental. *Experimental* designs are used in experiments to investigate cause and effect relationships. In contrast, *nonexperimental* designs are used in more naturalistic studies or in situations where the primary purpose is to describe the current status of the variables of interest. The latter designs are distinguished by the absence of manipulation by the researcher, with an emphasis on observation and measurement. Between these two broad categories are *quasi-experimental* designs which lack the randomization of *experimental* designs yet seek to address causal relations.

The adequacy of the research design to produce credible results, most notably to make causal inferences, is evaluated in terms of two primary types of validity: internal and external (Campbell & Stanley, 1963). *Internal validity* refers to the confidence that the specified causal agent is responsible for the observed effect on the dependent variable(s). *External validity* is the extent to which the causal conclusions can be generalized to different measures, populations, environments, and times. In addition, *statistical conclusion validity* is considered with internal validity and refers to the appropriate use of statistics. *Construct validity*, the ability to generalize the research operations to hypothetical constructs is a companion to external validity (Cook & Campbell, 1979).

Campbell and Stanley (1963) and Cook and Campbell (1979) produced the foundational works defining *quasi-experimental* design (see Desideratum 3) from which much of the literature on research design is extrapolated. Shadish, Cook, and Campbell (2002) revisited the initial works, providing greater attention to external validity, randomized designs, and specific design elements

Table 30.1 Desiderata for Research Design.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. The research design is foreshadowed and follows logically from the general problem statement and associated research questions.	I
2. The research problem is clearly articulated and researchable.	I
3. The research design is appropriate to address the research problem and is clearly articulated.	M
4. Variables are identified and operationalized; sampling, instrumentation, procedures, and data analysis are detailed.	M, R
5. The research design is internally consistent (e.g., the data analysis is consistent with the sampling procedures).	M, R, D
6. The design is faithfully executed or, if applicable, explanations of necessary deviations are provided.	I, M, R
7. Extraneous variability is considered and appropriately controlled.	M, R
8. Potential rival hypotheses are minimized. Threats to internal validity and statistical conclusion validity, and the adequacy of the counterfactual, are considered.	M, D
9. Conclusions as to what occurred within the research condition are appropriate to the design.	M, D
10. Generalizations, if any, are appropriate. External validity and construct validity are considered elements of the design.	M, D
11. The limitations of the design are articulated and appropriately addressed.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

to be used as a counterfactual rather than prescribed research designs. A white paper prepared for the American Educational Research Association (AERA) by Schneider, Carnoy, Kilpatrick, Schmidt, and Shavelson (2007) specifically addressed the issue of causal inferences using large-scale datasets, experimental, and nonexperimental designs. Texts by Keppel, Saufley, and Tokunaga (1998), Fraenkel and Wallen (2006), or Huck, Cormier, and Bounds (1974) provide more thorough introductory treatments, whereas the text by Keppel and Wickens (2004) presents more advanced coverage. Table 30.1 contains specific desiderata to guide reviewers and authors as they make decisions regarding quantitative research design.

1. The Research Design Is Foreshadowed

From within a quantitative social and behavioral science research framework, the discussion of research design usually begins with the methods or procedures used to conduct the study (e.g., the selection and/or assignment of participants, the operationalization of the variables, the procedures for data collection and analysis) and the subsequent implications for conclusions. However, the research design and methods utilized should not come as a surprise at the end of the Introduction, but rather should be an extension of the foundation that has been developed therein.

In describing research design for qualitative research, Maxwell (2005) identified five components that comprise his model for research design, much of which is applicable, yet has remained only implicit, for quantitative research design. The five interacting components that Maxwell identified include the goals, conceptual framework (which includes the theoretical framework and research literature), research question(s), methods, and validity. Although Maxwell included elements within these that are typically not appropriate to quantitative research (e.g., the inclusion of

personal experience within the conceptual framework), and he envisioned these elements dynamically interacting rather than the more sequential linear procedure of quantitative research, he did make explicit the need to evaluate the conclusions of a study within this larger context. It is in the Introduction that researchers should identify what variables will be attended to, which will be ignored, and which need to be controlled. The congruence of the Introduction, including its review of the literature, with the research design is necessary to evaluate the overall contribution of the study.

2. The Research Problem

It is impossible to evaluate the adequacy of a particular research design if there is no clearly articulated statement of the research problem. The research problem may be expressed in the form of research question(s) and/or hypotheses, and serves to formalize the research topic into an operational guide for the study, connecting the conceptual framework to the methods (Fraenkel & Wallen, 2006; Maxwell, 2005). The description of the research problem should identify the target population, the variables, and the nature of any anticipated relation between the variables and thereby focus the data collection and presage the data analysis. Hypotheses are not necessary, but are often stated when a specific prediction to be tested is made.

Terms used in the research problem must be defined in such a way that the questions are focused and testable. For example, “What is the best treatment for anxiety?” is not a testable question. Without defining “best” there is insufficient information to guide the study. Does “best” mean the most economical, the most consistent, or possibly the most permanent? The question also does not identify the population (e.g., children, teens, adults), or what types of treatment will be evaluated (e.g., psychotherapeutic, pharmacological, social behavioral), or what type of anxiety (e.g., self-report, clinically diagnosed, theoretically defined, physiologically measured). Without these further clarifications, it is not possible to assess whether or not the research design is appropriate to address the research problem.

Not only does the research problem suggest the appropriate research design, it also clarifies the specific type of data to be collected and thereby influences the data collection procedures. Questions can be classified as instrumentalist or realist (Cook & Campbell, 1979; Maxwell, 2005). *Instrumentalist* questions rely on the utilization of observable measures and require direct observation or measurement. *Realist* questions are about feelings, attitudes, or values that cannot be directly observed. The type of question, instrumentalist or realist, should connect the purpose of the study with the type of data collected. For example, if the purpose of the study is to provide information about teaching effectiveness, an instrumentalist question would be posed. It then would not be appropriate to collect the data using a survey within a survey research design to garner information from teachers as to their *perceived* effectiveness.

Designs developed for use with instrumentalist questions require greater inference and therefore might be more susceptible to bias. Yet as Tukey (1986) stated, and is often quoted, “Far better an approximate answer to the right question which is often vague, than an exact answer to the wrong question, which can always be made precise” (p. 407). Within the quantitative paradigm authors should clearly state their efforts to minimize bias and/or include appropriate caveats urging caution when interpreting and generalizing the results.

3. Articulation of the Research Design

The type of question(s), realist or instrumentalist (see Desideratum 2), will determine the type of data collected (e.g., self-report or performance). However, more fundamentally the research question(s)

will determine the appropriate type of research design. Questions about relations among variables or questions about the current status of variables can be answered with a nonexperimental design. Experimental designs dominate the discussion when it comes to questions about cause and effect. However, alternative designs have challenged the notion that experimental designs are the only type of design appropriate for causal inference.

The types of research design are distinguished by the degree to which the researcher is able to control the research environment. Four types of control are evaluated: (a) the researcher's ability to control the selection and/or assignment of participants to groups, (b) the manipulation of the independent variable(s), (c) how any dependent variables are measured, and (d) the timing of the measurement(s). The types of research designs vary significantly with regard to the type of control that the researcher is able to exert. Nonexperimental designs offer very little control, and experimental designs require more control.

Questions of cause and effect should be addressed using experimental designs. They are *experiments* in the sense that the researcher is able to control or deliberately manipulate conditions in order to observe the varying outcomes. As Shadish et al. stated:

Experiments require (1) variation in the treatment, (2) posttreatment measures of outcomes, (3) at least one unit on which observation is made, and (4) a mechanism for inferring what the outcome would have been without treatment—the so-called “counterfactual inference” against which we infer that the treatment produced an effect that otherwise would not have occurred.

(Shadish et al., 2002, p. xvii)

Within the category of experimental designs, *randomized* designs (sometimes referred to as *true experimental* designs), are distinguished by the researcher's ability to control the experimental conditions, most specifically the random assignment of participants to conditions. *Quasi-experimental* designs comprise a separate category because, although the researcher can manipulate the proposed causal variable and determine what, when, and who is measured, he/she lacks the freedom to randomly assign the experimental units or participants to the treatment conditions. Without this random assignment the researcher must be more circumspect when making causal inferences (Cook & Campbell, 1979; Shadish et al., 2002).

In addition to the randomized and quasi-experimental designs, methodologists have referred to *pre-experimental* or *pseudo-experimental* designs (Cook & Campbell, 1979; Huck et al., 1974) as forms of experimental designs. These designs are separated from quasi-experimental because of their lack of experimenter control and subsequent weaker claims of causality. It is imperative that researchers and reviewers attend to how the various types of control (and more specifically, the lack of control) can impact both the internal validity (see Desideratum 8) and external validity (see Desideratum 10).

Questions of cause and effect require a comparison. The ideal, but impossible, comparison is the *counterfactual* (Cook & Sinha, 2006). Whereas the experimenter is able to measure what occurs when a treatment is introduced, he/she cannot say what would have occurred to that individual had the treatment not been introduced—the counterfactual. Thus, any experiment requires an approximation of the true counterfactual: “The better the counterfactual’s approximation to the true counterfactual, the more confident causal conclusion will be” (Cook & Sinha, 2006, p. 551).

Specific experimental research designs are distinguished by how this counterfactual is constructed. Some designs use a control group and/or an alternate treatment, whereas others use a pretest measure to compare to the outcome measure. Some designs combine more than one approach (e.g., pre-test–post-test control group design) to improve the quality of the counterfactual. The logic of this approach is that the counterfactual represents what would be in the absence of

the treatment. Unfortunately, this belief cannot always be justified. For example, the use of a control group presumes that this group is identical to the treatment group in all ways except for the existence of the treatment. Similarly, the use of a pretest presumes that all else remains the same, except the exposure to the treatment. Clearly, these assumptions cannot always be defended and the researcher should provide as much evidence as is reasonable to support his/her claims of the adequacy of counterfactual that serves as the comparative. Additional variables should be tested to further support the argument of equivalence of a control group to which the participants have not been randomly assigned.

Nonexperimental designs are usually restricted to descriptive or associational research, where the main purpose is at most to provide evidence of relations between two or more variables. However, there are nonexperimental designs, or naturalistic designs, used to explore causal relationships (e.g., *ex post facto* or *causal-comparative*) (Fraenkel & Wallen, 2006) and studies utilizing advanced statistical modeling procedures. Due to the absence of researcher control over not only the assignment of participants, but also the manipulation of a hypothesized causal agent, causal conclusions can only be tenuously advanced from studies that use a causal-comparative design. Studies that utilize some form of statistical modeling (e.g., structural equation modeling; see Chapter 33, this volume) rely upon an *a priori* theory and stochastic assumptions to make causal claims. The use of hierarchical data or multilevel modeling share the same methodological issues as other designs in addition to some unique issues. In their review of 99 studies using “traditional” multilevel models, Dedrick et al. (2009) identified four broad issues: model development, data considerations, estimation procedures, and hypothesis testing and statistical significance. Insufficient information about the procedures, reliance upon the data and statistical procedures to guide decisions, and the large number of statistical tests make the results difficult to evaluate (see Chapter 22 of this volume and Dedrick et al., 2009, for discussion and specific recommendations for reporting and evaluating studies using multilevel methods).

Although the ability to determine causality has traditionally been randomization, a number of statistical procedures are used to equate groups when random assignment is not possible (or to evaluate the equivalency of groups formed via random assignment). Propensity score matching uses logistic regression to assign each individual a score from a list of potential confounding variables used as the predictors. It is presumed that by matching groups on the span of their propensity scores the influence of the nuisance factors have been eliminated and the groups are equivalent. However, this assumes that all of the confounding variables have been included in the determination of the propensity score; an assumption that warrants evaluation. A second issue emerges when the span of scores for the groups only marginally overlap. Only participants with scores within the overlap should be included, which may have the potential of increasing the probability of a Type II error due to the reduced sample size. This also introduces a potential bias in that individuals who remain may not be representative of their group. Individuals with the lowest scores will be eliminated from the group with the initial overall lower propensity scores, whereas the group with the higher span will eliminate individuals with the highest propensity scores. There exist a number of methods for equating the groups, which suggests a lack of consensus:

That there are so many alternative methods reflects the fact that no single approach is ideal and each has limitations. It is also disconcerting that, because each technique includes different subsets of people, it is quite possible to get different results depending on the choice.

(Steiner & Norman, 2012, p. 1380)

A *regression discontinuity design* is used when the probability of treatment is determined by where one scores on one or more naturally occurring or arbitrarily defined threshold(s) (e.g., remedial

course assignment below a set cut-score; eligibility determined by a specific age, geographic location, or income level). The assumption is made that those participants contiguous to the threshold are homogeneous and therefore the threshold arbitrarily serves to assign them to the control and treatment condition. The design is appropriate when the treatment is dichotomous with those on one side of the threshold becoming the control or comparison group and those on the other side comprising the treatment/experimental group. There are variations of the design (e.g., sharp versus fuzzy; Imbens & Lemieux, 2008), which have implications for the assumptions, data analysis, and interpretation of results. The design, when executed appropriately has strong internal validity. However, as with other quasi-experimental designs, the procedure does not ensure that the results are due to the treatment. Of particular concern is the possibility of selection bias due to an unidentified variable that is related to the threshold of the covariate or other changes at the cutoff (e.g., history threat). The validity of the regression discontinuity design is dependent upon the integrity of treatment assignment. The ability to manipulate an individual's score to alter the treatment assignment has serious implications as does compliance with treatment. The reviewer should evaluate the potential for misidentification, which if deemed extreme would invalidate the findings. Data analysis varies from relatively straight forward to very complex regression procedures, heuristically derived from visual inspection of the data. Attention to model specification is essential and evidence should be provided of efforts to assure unbiased and efficient estimates. Sensitivity tests should be conducted to assess the robustness of the results. Trochim (2006) presented an excellent introduction to the regression discontinuity design and analysis; Imbens and Lemieux (2008) provided coverage of designs and analysis with greater complexity.

The use of instrumental variables (IVs) occurs with both natural (observational) and randomized designs and with a number of statistical methods. By definition an IV must be correlated with the independent variable and uncorrelated with the error of the dependent variable; the effect on the dependent variable must be through the IV's relationship with the independent variable. Of critical importance is that use of an IV potentially changes the population to which the results can be generalized. The improvement in estimation is only for those who comply with the manipulation of the IV (i.e., the local average treatment effect, or LATE). In some cases, the research question of interest is addressed by the evaluation of the LATE, whereas for other questions the LATE represents only a subset (i.e., the compliers) of the population of interest. The ability to confidently determine the ratio of compliers to defiers has implications for the strength of the IV as well as the disparity between the LATE and the effect in the total population. If the treatment effects can be assumed homogenous then the LATE and the average treatment effect for the sample are the same. Sovey and Green (2011) suggested that although empirical evidence cannot be provided to assess the homogeneity, the researcher should present an argument based on the outcome of studies using different populations or IVs. An evaluation of the IV assumptions is required to determine the adequacy of the IV and whether an unbiased estimate of the effect has been achieved. In the randomized IV design (frequently an "encouragement design") the IV rather than the independent variable is randomly assigned. In nonexperimental studies IVs can be classified and evaluated on a gradient of plausible randomness (Dunning as cited by Sovey & Green, 2011). The randomness of the IV suggests that the IV is independent of the other predictors, however it does not guarantee that the IV effect on the criterion is only through the mediating factor. It is incumbent upon the researcher to defend that the IV satisfies both parts of the IV definition. Random assignment of the IV, a theoretical argument asserting the logic of the independence, as well as statistical tests to probe for potential correlations between the IV and other predictors can help to convince of the possibility that the IV is independent of unobserved variables related to the criterion. Whether or not the IV has a direct effect on the outcome should also be addressed by the researcher as he/she considers possible explanations. Weak instruments have the potential for bias and the strength of the IV should be tested. Although debated, the general guideline is that a single IV should have an

F greater or equal to 10. The effect of the IV must be monotonic (i.e., no one in the control condition receives the treatment). This again may depend on an argument rather than empirical evidence to justify why this assumption is met by the research design. The stable value treatment assumption (SUTVA) requires that the treatment or assignment of one unit not affect that of another. The use of more sophisticated sampling strategies (e.g., cluster sampling) may have implications for the satisfaction of this assumption. Four final considerations for the IV design: (a) the use of IV requires large sample sizes, (b) the estimated effect may be biased if the IV is not dichotomous, (c) the various statistical procedures also have assumptions that must be evaluated, and (d) different analyses models may result in different results.

Despite the statistical sophistication, methodologists remain divided on whether nonexperimental designs can provide convincing evidence that warrant claims of causality (see, for example, Shaffer, 1992).

Researchers who rely upon extant databases should be attentive to the quality of the original research design and how the design decisions impacted the data collected. For example, large datasets frequently result from sampling strategies that have implications for how the data should be evaluated (e.g., weighting). Researchers using existing data, including those performing statistical modeling, should (a) disclose information relevant to how the data were obtained, (b) provide sufficient detail of the *a priori* theory or theories, (c) faithfully execute the chosen statistical procedure after adequately addressing associated underlying assumptions, and (d) acknowledge the limitations of the study to make claims of causality.

The selection of the research design should consider the research problem within the larger context of the research topic. Careful consideration should be given to whether a longitudinal within-subjects design (see Chapter 2, this volume) or a cross-sectional between-subjects design (see Chapter 1, this volume) would be better suited to address the research problem. For example, either design can answer the question of whether or not there is a difference in performance on some defined measure of knowledge of teenagers and septuagenarians. However, if the hypothetical construct being measured is long term memory, the longitudinal design will enable greater confidence that the difference in test performance is due to memory rather than learning. If, however, the hypothetical construct represented by the test performance is learning, the less time and cost consuming, cross-sectional between subjects design would be adequate and consistent with research in this field.

Once determined, the research design helps authors to coordinate how participants are selected, how variables are manipulated, how data are collected and analyzed, and how extraneous variability is controlled. Discussion of specific designs can be found in Cook and Campbell (1979), Huck et al. (1974), Shadish et al. (2002), Trochim (2006), or Creswell (2005). Each element of the research design should be described with sufficient detail that the study can be replicated. All variables (i.e., independent, dependent, moderator, mediator, or control) should be defined, and the measurement should be congruent with the presentation in the Introduction. The type of design (i.e., experimental, quasi-experimental, or nonexperimental) or the specific design (e.g., groups \times trials mixed between-within design, or nonequivalent control group design) should be stated. Adherence to a specific design is not required and the inclusion of additional procedures to control extraneous variability is encouraged (e.g., the inclusion of a pretest or a control group). In their follow-up text to Campbell and Stanley (1963) and Cook and Campbell (1979), Shadish et al. (2002) emphasized the value of design elements as counterfactuals rather than designs per se. In essence the design is constructed rather than selected from a prescribed list. The inclusion of each design element should be evaluated in terms of the potential impact on both internal and external validity (see Desiderata 8 and 10).

It is not uncommon for the researcher to omit any explicit reference as to what research design or design elements are used. Yet as Maxwell (1996, p. 3) noted, “Research design is like a philosophy of life; no one is without one, but some people are more aware of theirs and thus able to make more informed and consistent decisions.” When design elements have not been explicated, the degree to which the researcher has made conscious design decisions is unknown. In this case, not only must the reviewer be vigilant in evaluating the credibility of what is reported, but he/she must also try to reconstruct the design that was used by what is reported in the Methods and Results sections. Without the aid of the author to define the counterfactuals used, the reviewer and reader are left to not only evaluate their effectiveness, but to also identify what they are. This is essential to determining whether the research design can support the stated conclusions.

4. Specific Design Elements

The first element of the research design is a description of the participants. The selection of participants should be consistent with the identified design. The type of design will determine first whether group assignment is necessary, and second, if so does assignment precede or follow selection. If a sample is used, the sampling frame and the population should be identified. The sampling procedure should be specified and there should be a justification of the sample size (see Chapter 35 this volume). An appropriate sample, in size and composition (i.e., representativeness), is foundational to the conclusions of the study (see Desideratum 9). In addition, how participants are assigned to treatment conditions (if appropriate) is important to the determination of the strength of any inference of causality (see Desideratum 8). Not only is it important to report what definition or instrument is used to select or assign participants, it is also important to report the reliability, validity, and cut scores of that instrument. This provides confidence that the participants met the criterion established and allows comparison to previous research. For example, “extraverts” as defined by the Eysenck Personality Inventory (EPI) are not the same as “extraverts” defined by the Myers–Briggs Type Indicator. Defining extraverts as those scoring above the sample mean on the EPI may not be the same as extraverts defined as those scoring above a normed score. The method by which participants are placed into groups (i.e., no groups, randomly assigned, or pre-existing groups) is essential to the type of research design being used and therefore to the conclusions that can be drawn (see Desiderata 8 and 9).

An integral element to the integrity of any study is the reliability and validity of the instrument(s) used to collect the data (see Chapter 29, this volume, for specifics on validity and reliability assessments). Not only is it necessary to provide evidence of the appropriate types of reliability and validity that have been established, but to make the case for why the author would expect that this evidence would apply to his/her use of the instrument. Citing extensive previous use is not sufficient evidence of reliability or validity.

The experimental and/or data collection procedures comprise the next element of the research design. There should be a detailed description of any experimental conditions, including any control conditions, if a treatment is introduced. This should include precise details of time intervals—duration of exposure to the treatment as well as time lapse between exposure and data collection, dosages, equipment settings, and research personnel. How and when the data are collected should be clearly described, making special note if the timing or the mode of collection could affect the response. For survey research designs this should include the number of reminder contacts, the timing of the reminders, and the mode of contact.

The final element of the research design before the discussion of study results is the presentation of the data management and data analysis. Data reduction and transformations, including

the treatment of missing values should be articulated, highlighting any deviations from standard procedures. The data analysis should explicitly address demographic data that are useful for the discussion of appropriate generalizations (see Desideratum 10) or the equivalency of groups (see Desideratum 3). Data from instruments with total or scale scores should be analyzed for internal consistency reliability and compared to previous uses of the instrument. The specific test(s) used to address each research question and/or hypothesis should be named, including any information necessary for the reader to determine the appropriateness of the test or decision (e.g., degrees of freedom, alpha level, p values). *Post hoc* tests for the interpretation of omnibus test results (e.g., *post hoc* comparisons following an ANOVA) must be included and should be identified by name. When using samples, there should be a test at every point of decision. Look for words of comparison—most, greater, fewer, and verify that the appropriate statistical test has been conducted. When multiple tests are reported, consider the potential for an inflated Type I error rate. Evidence should be presented to confirm that the assumptions of statistical tests were met or that appropriate adjustments were made.

5. Internal Consistency of Research Design

A research design that lacks internal coherence creates problems for the interpretation of the results. This problem emerges particularly when the research design has not been explicated. Beginning with the Introduction, which should establish the need for the study and what precisely will be studied, through the statement of the research problem and research question(s), the way the sample is selected, the independent variable(s) are manipulated, the data collected, and how the data are analyzed, each design element should logically follow. If the researcher claims that a randomized design is used, it then follows that participants must be randomly assigned to the experimental conditions. If the research question is about differences between groups, the sampling plan must be such that ensures sufficient representation in each group and not left to random selection. The most blatant example of inconsistency is when the statistical analysis is not appropriate for the type of research question or how the data were collected (e.g., using a test of correlation to answer questions of cause and effect, especially when no temporal order has been established). Similarly, if participants are selected because they represent the two extremes of a grouping variable, it would be inappropriate to use correlation to evaluate the relationship between the two variables.

6. Design Execution

Details that researchers present in the Methods section must be consistent with what they intended at the outset of the research, as expressed in the Introduction. The procedures detailed in the Methods section should be evaluated to verify that they were faithfully executed. Small departures from the original design are often unavoidable—even anticipated, however, they require explanation.

A common problem is that the number of anticipated observations is not equal to the sample size upon which the conclusions are based, likely reducing the desired power level (see Chapter 35 this volume). This issue is particularly prevalent in survey research. Even in studies where researchers over-sample in an effort to achieve the desired sample size, the total response rate is often less than desired; the response rate for an individual survey item (i.e., missing response) may be considerably lower (Jackson, 2002). Often, studies are designed with equal or proportional sample size in each cell, yet frequently when the results are reported the cells are uneven. This has implications for how missing data are treated, for statistical assumptions, for the power of the test, as well as for other design implications. It is therefore incumbent upon the researcher to account for missing values and evaluate the implications for the design. One consideration is whether the nonresponses

adversely affect the representativeness of the sample. More specifically, consideration must be given to whether missing values represent a threat to internal validity (see Desideratum 8). The disproportionate loss of participants from one treatment condition might suggest a threat to internal validity (Cook & Campbell, 1979). This is not only true of quasi-experimental designs but also of randomized designs, which by virtue of random assignment of participants are protected from most other threats.

Reviewers of manuscripts should be vigilant for evidence suggesting that procedures were counter to stated claims. For example, if a researcher claims that participants were randomly assigned, but then later suggests that the treatment was assigned to pre-existing groups, this changes the design from a randomized design to a quasi-experimental design with all the attendant issues that must be addressed.

Valid conclusions about a causal relation between treatment and outcome are dependent upon the treatment (and control) condition(s) being faithfully delivered and the dependent variable reliably measured. There should be evidence that the researcher (or whoever is providing the treatment) has been trained and dependably delivers the specified treatment. Evidence in the form of manipulation checks should be provided to verify that experimental manipulations were effective. For example, experiments that rely on deception require that the participants are indeed deceived, and a well-designed study will provide evidence to this effect. In addition to ensuring that the experimental conditions are consistent with what is reported there should be evidence that the researcher has sufficient training to collect the data (e.g., training for interviewers, inter-rater reliability).

In addition to considering whether the researcher has delivered treatment successfully, the researcher and the reviewer should consider the plausibility of participant noncompliance with treatment. Drug trials are dependent upon participants actually consuming the prescribed dose; training is dependent upon participant attendance. Without evidence of participant compliance there is insufficient evidence that the research design has been implemented.

7. Control of Extraneous Variability

Without appropriate control of extraneous variability it can be difficult to isolate and observe the effect(s) of the hypothesized causal variable(s) on the dependent variable(s). Control of extraneous variability is therefore one of the primary functions of a research design. Rigorous adherence to carefully designed research procedures can help to minimize the effect of unintended influences. However, the design element that has the greatest impact on the control of extraneous variability is the selection and/or assignment of participants. Random assignment to treatment conditions is the principal means by which a research design avoids the systematic influence of unintended variables. The advantage of this method is that it controls for the influence of a number of variables, even those unidentified. However, it alone might be insufficient if an extraneous variable has a stronger effect than the causal variable that is being considered. Manuscript reviewers should note any mention of one or more variables in the Introduction (or from previous content knowledge) that is known to have a strong relation to any of the dependent variables, and ensure that its influence is considered in the research design chosen by the study's authors. In fact, an alternative research design might have been more appropriate. For example, the effect of an extraneous variable might be controlled by including it as an additional variable in the design (i.e., randomized block design) or by restricting the population of the study to only one level of the extraneous variable (e.g., only include women in the study).

Matching is a procedure whereby participants are paired on their scores for a specific variable(s) and then each member of the pair is assigned to a different treatment condition. The intent of using this procedure is to equate the groups in terms of this specific variable, a variable that is believed

to influence the dependent variable. This procedure should be used judiciously. Although it might equate the groups on that one specific variable, matching interferes with the ability to randomly assign participants, and thereby forfeits the benefit of randomization. The implications for the data analysis also must be considered. The matched pairs cannot be treated as independent observations and the data analysis must reflect this. The use of these procedures must be considered within the larger context of the overall design to ensure that their use is reflected in other design decisions (e.g., data analysis) and conclusions.

Sometimes, statistical procedures can also be used to control extraneous variability. The use of covariates can help adjust the scores on the dependent variable before testing for group differences if the extraneous variable is measured as a continuous variable. Propensity scores from a logistic regression (see Chapters 16 and 28, this volume) might also be used to evaluate the equivalence of treatment groups, improve matching, and/or be used as a covariate (Pasta, 2000). Although the use of propensity scores is a means of controlling for the effect of more than one extraneous variable, it is still limited to the control of only those variables that are identified and quantitatively measured. Rather than attempt to improve the equivalence of groups, Rosenbaum (1991) suggested a procedure (hidden bias sensitivity analysis) to assess how much bias would be necessary between the treatment and control groups for bias to be a viable alternative explanation for the treatment outcome. Shadish et al. (2002) warned that the use of these, or other advanced statistical procedures, is not a substitute for good design. Where possible, extraneous variability should be controlled by the research design, and then if appropriate augmented by available statistical procedures.

8. Internal Validity and Statistical Conclusion Validity

The careful construction and faithful execution of the research design provides the foundation for the research conclusions. Each element of the design relates to the validity of the study. Research conducted to test causal relations relies on the adequacy of the constructed counterfactual to represent the true counterfactual (see Desideratum 3). The adequacy is evaluated in terms of the ability to rule out rival hypotheses or alternative explanations for the outcome. In 1957, Campbell first coined the term *internal validity*, which was further elaborated by Campbell and Stanley (1963) as the confidence that the identified causal variable is responsible for the observed effect on the dependent variable and not due to other factors. They identified a list of threats to internal validity, which should be considered when constructing the design as well as when evaluating the conclusions. The list of threats to internal validity, with some modifications, can be found in most research design textbooks (also see Shadish et al., 2002; Shadish & Luellen, 2006; for discussions on threats relevant to specific designs see Cook & Campbell, 1979; Huck et al., 1974).

Threats to internal validity are usually discussed in terms of quasi-experimental designs because they result from the inability to randomly assign participants to treatment conditions. That is, random assignment reasonably protects the study from most threats to internal validity; however, such threats should be considered for any study that seeks to make causal inferences, with or without random assignment. Some threats (e.g., mortality or attrition, the disproportional loss of participants from one condition) occur after the assignment to experimental condition or as a result of something that occurs during treatment delivery, which thereby jeopardize causal interpretations of even a randomized design.

The potential of a threat to internal validity in and of itself is insufficient to dismiss a researcher's claims of causality. When evaluating the potential threats, Shadish and Luellen (2006, p. 541) advocated the consideration of three questions: "(a) How would the threat apply in this case? (b) Is the threat plausible rather than just possible? and (c) Does the threat operate in the same direction as the observed effect so that it could partially or totally explain that effect?" If it can be conceived how a

specific threat would offer a rival hypothesis, which is probable—not just possible, and explains the direction of the outcome, only then would the internal validity be challenged. The careful researcher will consider these threats in the design of the study, anticipating those with potential relevance. If considered prior to the execution of the study, it may be possible to alter a design element(s) to avoid a potential threat, or additional data may be collected to provide evidence to argue against a threat's explanatory ability (see Desideratum 3).

Cook and Campbell (1979) further refined the discussion of internal validity by introducing *statistical conclusion validity* as a distinct form of validity related to internal validity. Statistical conclusion validity refers to the “appropriate use of statistics to infer whether the presumed independent and dependent variables covary. Internal validity referred to whether their covariation resulted from a causal relationship” (Shadish et al., 2002, p. 37). Threats to statistical conclusion validity provide reasons why the researcher might be wrong about (a) whether a relationship exists and (b) the strength of the relationship. A list of threats to statistical conclusion validity can be found in Shadish et al. (2002, p. 45). Attention to statistical power, assumptions of the statistical tests, inflated Type I error, and effect size, as well as issues related to the measurement and sampling, fall within the purview of statistical conclusion validity.

Generally, it is not appropriate to refer to the internal validity of nonexperimental designs, with one exception: specific designs that are being used to make causal inferences (e.g., causal comparative). However, the validity of conclusions reached still requires evaluation. Each design decision affects the validity, with decisions regarding the appropriate sampling, instrumentation, and statistical analysis of particular importance for the nonexperimental design. The sample size and representativeness of the population, the reliability and validity of measurement, and the appropriate statistical analysis are key to the conclusions of a nonexperimental study.

Authors and reviewers must keep in mind that “Validity is a property of inferences. It is *not* a property of designs or methods, for the same design may contribute to more or less valid inferences under different circumstances” (Shadish et al., 2002, p. 34). Executing a prescribed design does not guarantee valid inferences, nor does the rigid adherence to a checklist of potential threats to validity. Neither are adequate substitutes for the researchers’ sound logic.

9. Conclusions Are Appropriate

Miles and Huberman (1994) noted that design decisions both support and constrain the conclusions of research. Just as the genesis of the research design is before the Methods section, its influence extends beyond the Results. Researchers are responsible for presenting conclusions that are consistent with and appropriate to the design. The adage “correlation is not causation” is just one example for the necessity to ensure that claims in the Discussion do not exceed what the research can support. Design decisions, such as the decision to control extraneous variability to only one level of an extraneous variable (e.g., women only) restrict the conclusions to only that group.

Careful articulation of the research design elements, with attention to potential threats to internal and statistical conclusion validity (see Desideratum 8), prepares the researcher to present the conclusions within the context of the existing literature. Causal claims should not be made without ruling out threats to internal validity. With a nonexperimental design utilizing only descriptive statistics to report the findings from a sample, it is inappropriate to make comparisons between groups (e.g., “women scored higher than men”). There must be a test at the point of decision.

Without appropriate supporting evidence it is inappropriate to draw conclusions from statistical nonsignificance. For example, when testing for mean differences between treatment populations, nonsignificance should not be interpreted to imply that there is no difference between the populations’ means or that the population means are therefore equal. Statistical nonsignificance means

that the researcher has failed to show a difference of sufficient magnitude that cannot be reasonably explained by chance alone. That the population means are equal is only one possible explanation. It is also possible that the sample size was insufficient or the measurement not sensitive enough to detect true differences.

The tendency to overstate findings is not limited to misrepresenting statistical conclusions or failing to recognize threats to internal validity, but also includes making claims beyond what was studied. For example, if a study using a survey to measure the level of teacher satisfaction shows that 65% of teachers report being *slightly dissatisfied* with teaching, it is inappropriate for the researcher to conclude that his/her study found that teachers will be leaving their schools, or that teachers should be paid higher salaries. The author should be diligent to ensure that recommendations from the study are not presented in a manner that they can be construed as findings.

10. External Validity and Construct Validity

Technically, the *external validity* of a research design refers to the degree to which a study's observed *causal* relations are generalizable; that is, it helps characterize "to what populations, settings, treatment variables, and measurement variables can this effect be generalized" (Campbell & Stanley, 1963, p. 5). Internal and external validity are considered to be complementary: whereas the former addresses the question of what can be inferred about cause and effect from this instance, the latter assesses the degree to which the causal findings can be generalized. Frequently what will increase internal validity may decrease external validity and vice versa. In their 1979 work, Cook and Campbell extended their dichotomous discussion of validity into the typology that comprised internal, statistical conclusion, external, and construct validity. Whereas internal and statistical validity (see Desideratum 8) are relevant to the inferences that derive from the specifics of the study procedures, construct and external validity relate to whether the inferences can be extended beyond the current situation. *Construct validity* generalizations refer to "inferences about the constructs that research operations represent" and external validity generalizations are "inferences about whether the causal relationship holds over variation in persons (or more generally: units), settings, treatment, and measurement variables" (Shadish et al., 2002, p. 20). From these definitions it becomes apparent that with nonexperimental designs that are used to describe the current status or noncausal relations between variables, it is inappropriate to discuss external validity. Instead, construct validity is the more appropriate consideration. Thus, nonexperimental designs that are not used to test causality should be evaluated for construct validity, and nonexperimental designs that are used to evaluate causal relations and experimental designs should be evaluated for both construct and external validity.

Construct validity is inherent in social and behavioral research and as an issue is twofold: definition and measurement. Every construct has multiple facets or features, with some being more central than others. Thus, defining the construct requires identifying multiple components, with the core being those features to which there is the greatest agreement. Once defined, the question becomes one of how to represent the construct, and more specifically how to measure it. Determining how multi-faceted constructs can be reduced to a manageable size, yet still represent the higher order construct, is the dilemma of construct validity. Each study uses a limited set of conditions in terms of the population, the treatment, the setting, and the outcome; from which the desire is to make statements about the higher order construct. Each element of the research design should be evaluated for the construct(s) it represents. Researchers tend to focus on only the treatment variable, if there is one, and the outcome measure. Clearly, discussions limited to the construct validity of the outcome measure are insufficient as they address only one of the constructs in the study. How will the sample selected reflect the larger construct that it represents? For example, how does a sample consisting of students two grades below reading level represent a population of "students at risk"?

How does conducting the study in the laboratory represent the larger construct of the settings where the conclusions would apply? These questions need to be considered in addition to the more obvious examination of how the treatment and outcome constructs are operationalized. There is no one-to-one correspondence of the operationalization of the study and the constructs; the question is: *How great is the disparity?* A list and further discussion of potential threats to construct validity is presented in Shadish et al. (2002).

External validity refers specifically to whether or not observed *causal* relations can be extended across individuals, settings, treatments, and/or outcome measures. The use of probability sampling is the foundation for external validity. Probability sampling requires that each item in the domain has a nonzero chance of being randomly selected. This condition is infrequently met when sampling participants for a study, much less when sampling from the domains that describe the other elements of an experiment (i.e., the setting, treatment conditions, outcome measures). Shadish et al. (2002) explicated a more heuristic approach to determine causal generalizations. They proposed five principles for consideration: surface similarity, ruling out irrelevancies, making discriminations, interpolation and extrapolation, and causal explanation. Too frequently external validity is only discussed in terms of generalizing to populations, either those internal or external to the study, and the ability to generalize to the other elements receives scant attention. With a design seeking to establish evidence of a causal relationship, the researcher and reviewer should examine the degree to which the design elements that are included represent a random sampling of the construct domain, be it the population, setting, treatment, or outcome. In the absence of random sampling, the principles described by Shadish et al. (2002) provide a systematic means to evaluate external validity. Shadish et al. also present a list of common threats to external validity.

11. Design Limitations

The diligent researcher will acknowledge weaknesses in the research design and present the implications of the shortcomings. For example, by recognizing in advance that the use of pre-existing groups compromises the internal validity (see Desideratum 8), the researcher has the opportunity to offer explanations, possibly even statistical evidence (see Desideratum 7), to argue the equivalence of the groups prior to the introduction of the treatment. By ignoring any reference to this design decision, the reviewer and reader are left to decide whether the potential pre-existing group differences are sufficient to explain the outcome. More significantly, this can create a lack of confidence. The question becomes: *If the researcher does not know enough to discuss the implications of the use of pre-existing groups, what other relevant information might he/she not recognize the necessity to reveal?* Does the researcher understand enough about the research design to adequately convey the information necessary for the reader to make an independent decision as to the appropriateness of the conclusions?

In theory, many of the weaknesses can be avoided by assiduous attention to the research design, yet this is not always the case. Weaknesses result from a lack of feasible alternatives, unforeseen occurrences during the study, and/or from poor research design. The credibility of the researcher is enhanced if he/she is able to eliminate him/herself from the latter category by anticipating and addressing criticism from the knowledgeable reviewer or reader.

References

- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago, IL: Rand McNally.
 Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

- Cook, T. D., & Sinha, V. (2006). Randomized experiments in educational research. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 551–565). Mahwah, NJ: Erlbaum.
- Creswell, J. (2005). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, NJ: Prentice Hall.
- Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, 79, 69–102.
- Fraenkel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research in education* (6th ed.). New York: McGraw Hill.
- Huck, S. W., Cormier, W., & Bounds, W. G. (1974). *Reading statistics and research*. New York: Harper & Row.
- Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142, 615–635.
- Jackson, G. B. (2002). *Sampling for social science research and evaluations*. Retrieved from www.gwu.edu/~gjackson/281_Sampling.PDF
- Keppel, G., Saufley, W. H., Jr., & Tokunaga, H. (1998). *Introduction to design and analysis* (2nd ed.). New York: Freeman.
- Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). New York: Pearson Prentice Hall.
- Maxwell, J. A. (1996). *Qualitative research design: An interactive approach*. Thousand Oaks, CA: Sage.
- Maxwell, J. A. (2005). *Qualitative research design: An interactive approach* (2nd ed.). Thousand Oaks, CA: Sage.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage.
- Pasta, D. J. (2000). Using propensity scores to adjust for group differences: Examples comparing alternative surgical methods. In *Proceedings of the twenty-fifth annual SAS Users Group International Conference* (paper 261–25). Cary, NC: SAS Institute.
- Rosenbaum, P. R. (1991). Discussing hidden bias in observational studies. *Annals of Internal Medicine*, 115, 901–905.
- Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). Estimating causal effects using experimental and observational designs: A think tank white paper. Washington, DC: Prepared under the auspices of the American Educational Research Association Grants Program.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Shadish, W. R., & Luellen, J. K. (2006). Quasi-experimental design. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research*. (pp. 539–550). Mahwah, NJ: Erlbaum.
- Shaffer, J. P. (Ed.). (1992). *The role of models in nonexperimental social science: Two debates*. Washington, DC: American Educational Research Association and American Statistical Association.
- Sovey, A. J., & Green, D. P. (2011). Instrumental variables estimation in political science: A readers' guide. *American Journal of Political Science*, 55, 188–200.
- Steiner, D. L., & Norman, G. R. (2012). The pros and cons of propensity scores. *Chest*, 142, 1380–1382.
- Trochim, W. M. K. (2006). Design. Retrieved from www.socialresearchmethods.net/kb/design.php
- Tukey, J. W. (1986). *The collected works of John W. Tukey: Philosophy and principles of data analysis 1949–1964, Volume III* (Ed. L. V. Jones). Boca Raton, FL: CRC Press.

31

Single-Subject Design and Analysis

Andrew L. Egel, Christine H. Barthold, Jennifer L. Kouo, and Faye S. Maajeeny

Single-subject research is a form of rigorous investigation in which the individual is the unit of analysis. Individual variability of each participant is measured as opposed to the mean performance of groups. This allows for the examination of individual responding of participants during and following an intervention using previous performance (i.e., baseline measures) as a control (Sidman, 1960). Variability and experimental control are evaluated through visual inspection of graphed data (Skinner, 1938) and confirmed through independent and systematic replication both within and across research studies. Therefore, it is a misnomer that single-subject designs have an N of 1; in fact, most studies have three or more participants.

Single-subject designs can be used to study any construct defined so that it can be observed and measured over time, such as methods for effective teaching in higher education (Saville, Zinn, Neef, Van Norman, & Ferreri, 2006), teaching math skills to at-risk children (Mayfield & Vollmer, 2007), or psychological phenomena such as severe phobias (Jones & Friman, 1999). The results of single-subject designs are often the identification of effective interventions for heterogeneous populations where random assignment can be compromised or when information is needed about variable performance within groups (e.g., students with learning disabilities or autism). Kennedy (2005) suggested that single-subject designs can be used to demonstrate the effectiveness of interventions, to compare two or more interventions, and to complete parametric and component analyses. Some useful resources for those new to single-subject designs are Gast and Ledford (2014), Kratochwill and Levin (2014), Kennedy (2005), Bailey and Burch (2002), and Horner et al. (2005).

1. Participants and Setting

Careful description of the participants' characteristics and where the intervention is carried out are of high importance in any type of study. However, more emphasis is placed on participants and setting in single-subject research because of the importance placed on understanding individual variability. Several factors have been identified in the literature that influence selection of settings and participants.

Table 31.1 Desiderata for Single-Subject Design and Analysis.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. For purposes of future replication, participants and setting(s) are described in detail.	M
2. The experimental procedures utilized are described in enough detail to allow for replication and control for threats to internal validity.	M
3. The experimental design is selected based on the specific research question(s) investigated.	M
4. Within-experiment replications are consistently applied and sufficient to show experimental control.	M
5. Measurements of independent and dependent variables are operationally defined, and include direct observations of the phenomenon to be studied.	M
6. Inter-observer agreement coefficients are presented, together with associated formulae used (dependent on the dimension of behavior recorded). The inter-observer agreement is sufficient to indicate reliable interpretation of operational definitions.	M, R
7. Data are presented graphically to allow for visual inspection and decisions on when to alter variables are based on level, trend, consistency across phases, variability, and overlap.	R
8. If statistical analyses and randomization are used, procedures are in agreement with best practices.	R

* I = Introduction, M = Methods, R = Results, D = Discussion.

Important factors in choice of setting. Settings must have face validity relative to the settings in which the intervention is likely to be adopted. For example, research designed to increase social interactions between children with autism spectrum disorders and typical children should be conducted in schools if teachers and classroom staff are the intended audiences. Conducting the same study in a child's home would not have face validity.

Settings should also be stable and flexible throughout the study. Settings with rapidly changing schedules or ongoing issues (e.g., a classroom with substitute teachers) present confounding variables that are not easily controlled. Settings with little flexibility in scheduling and/or choices of intervention might make data collection difficult or impossible (Bailey & Burch, 2002).

Factors that affect the selection of participants. Care must be taken to provide a detailed description of the participant, including characteristics that might be related to the study's outcomes (e.g., communication ability, psychiatric diagnoses, and age). Complete subject descriptions can, at times, provide possible explanations for failure in replication and lead to a greater understanding of the generality of results (Kazdin, 1981).

Bailey and Burch (2002) identified several characteristics that will affect selection of participants. They noted that participants must reflect the dimensions of the general population that are being addressed in the study. For example, if the investigation is designed to increase interview skills, then the participants should be individuals who are struggling with those behaviors. Participants need to be readily available over the course of the study, and demonstrate stability with respect to health, cooperation, and attendance.

2. Experimental Procedures

Experimental procedures should be selected so that functional relations between the dependent and independent variable can be demonstrated, threats to internal validity controlled, and social validity (i.e., the extent to which the change in the target behavior improves the participant's quality of life) maximized. Most research studies using single-subject designs will have, at a minimum, a baseline and intervention phase. *Baseline* refers to the experimental condition that precedes the intervention phase. Baselines allow for a contextual evaluation of the effects of the independent variable. Although baseline is often thought of as a control condition, another experimental condition might serve as baseline as well. For example, if a person with heart disease is participating in a study to determine the effects of an experimental medication, baseline might be taken on symptoms while the participant is on a more well-established medication. All aspects of both the baseline and intervention conditions, including the exact materials used, any instructions provided, and levels of feedback for correct/incorrect responding, must be described in enough detail to permit replication.

It should also be clear that researchers designed the experiment to control for threats to internal validity. Like all research designs, the validity of single-subject designs can be threatened by variables such as history and maturation. Conversely, threats such as regression to the mean, participant selection bias, and selective attrition are not considered threats to internal validity given the single-subject nature of the designs (Kennedy, 2005). However, there are some threats to internal validity that are of particular concern to the single-subject researcher, as described below.

Testing, or repeated exposure, is of concern because the frequent number of measurements characteristic of single-subject designs could result in the participant learning the response(s) independent of intervention. Testing effects can be minimized by spacing out observations and/or choosing an experimental design in which the number of data points needed is minimized.

Multiple treatment interference occurs when participants receive more than one intervention in a condition and it is a serious threat to the internal validity of a single-subject design. Multiple treatment interference does not allow a researcher to determine which of the interventions, alone or in combination, were responsible for changes in behavior.

Sequence effects can also influence interpretation of data collected within a single-subject design. They occur when a variable is introduced or removed in a particular order and effect responding in subsequent conditions. Sequencing the introduction and removal of variables is crucial in single-subject designs and a researcher must be sensitive to the possibility of their occurrence. Counterbalancing the sequence of conditions across participants is one way to control for sequence effects.

3. Experimental Design / Research Question Correspondence

There are several single-subject designs that can be used to analyze the effects of an intervention. Unlike group designs, single-subject designs use the individual as their unit of analysis. As such, attrition of even one participant can be detrimental to a study (e.g., Christ, 2007). The design a researcher selects will depend on the specific question asked as well as the resources that are available at the time. In each of the designs discussed below the individual serves as his or her own control and the experimenter replicates the effect(s) of the independent variable in order to establish experimental control.

The *reversal design* (Baer, Wolf, & Risley, 1968) has, historically, been one of the most frequently used single-subject design. This design requires that the experimenter implement both the baseline and treatment phases multiple times in order to demonstrate experimental control. An A-B-A-B reversal design is typically used because it allows researchers to replicate both the baseline ("A") and intervention ("B") conditions.¹ Experimental control is established when the data patterns in each condition co-vary with the introduction and removal of an independent variable. The effects must

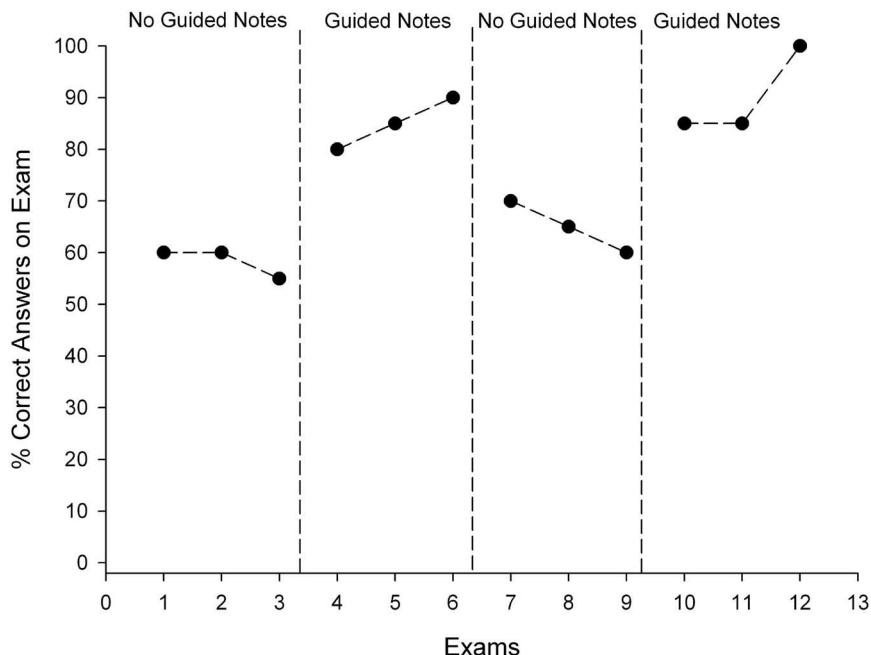


Figure 31.1 Sample ABAB Graph.

be replicated across each condition. An A-B-A-B reversal design is illustrated in Figure 31.1 as well as in the study by Ahearn, Clark, MacDonald, and Chung (2007). Ahearn et al. used a withdrawal design to evaluate the effects of response interruption and redirection (RIRD) on the occurrence of vocal stereotypy in four children with autism spectrum disorder (ASDs). The authors found substantial decreases in vocal stereotypy for each participant when RIRD was implemented.

Although reversal designs can be used to demonstrate the effects of an independent variable, there are circumstances where it would be inappropriate to use such a design: irreversibility of a behavior and when reversing the behavior puts the participant at risk of injury. In these circumstances, other single-subject designs would be more appropriate. A *multiple baseline design* is one of the design alternatives. Multiple baseline designs require the concurrent collection of three or more baselines (across participants, behaviors, or settings; cf. Kratochwill et al., 2010).

When responding is consistent across baselines, the intervention is introduced systematically to one baseline at a time. Experimental control is demonstrated when the behavior changes only when the intervention is implemented and the effects of intervention are replicated across participants, behaviors, or settings. An example of a multiple baseline design across participants is presented in Figure 31.2 and can be seen in the investigation by Briere, Simonsen, Sugai, and Myers (2015). Briere et al. examined the extent to which a within school consultation intervention could be used to increase specific rates of social praise by new teachers. The authors introduced the intervention successively across participants, in a multiple baseline fashion. The results showed that all three new teachers increased their rate of social praise after experiencing the intervention.

Horner and Baer (1978) introduced a variation of the multiple baseline design for situations in which baseline responding will be zero or when extended baselines can result in high rates of problematic behavior (e.g., tantrums, noncompliance). The *multiple probe design* requires that baseline sessions be collected intermittently rather than continuously as in the multiple baseline design.

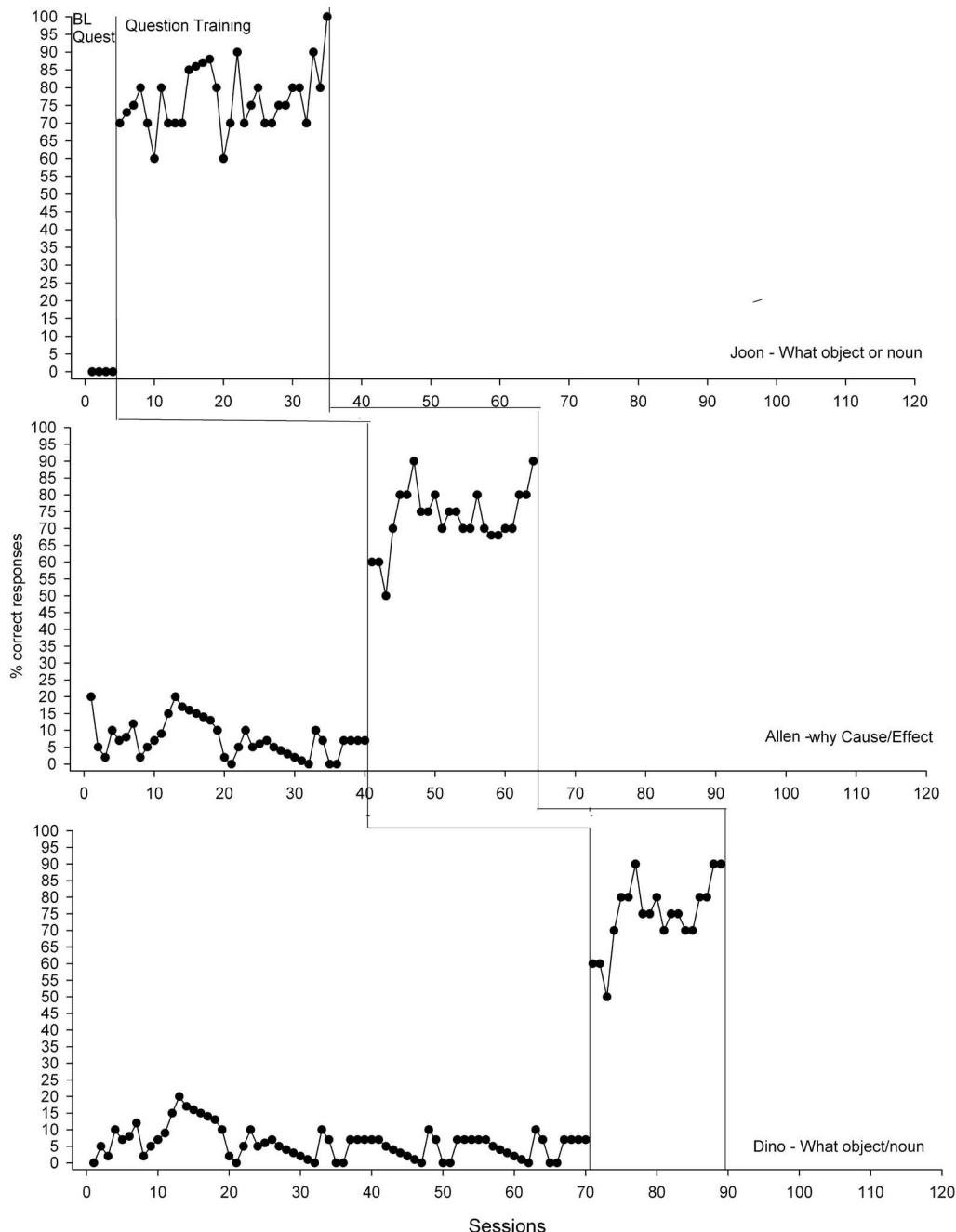


Figure 31.2 Sample Multiple Baseline Graph.

According to Horner and Baer (1978), implementation of a multiple probe design requires that an initial baseline probe be collected across each behavior, participant, or setting. One to two additional probes are conducted on the first tier of the multiple baseline design. Intervention is subsequently implemented on the first tier, while no data are collected on any of the remaining tiers. Once the data in the first tier show an effect or reach criterion, an additional probe is implemented

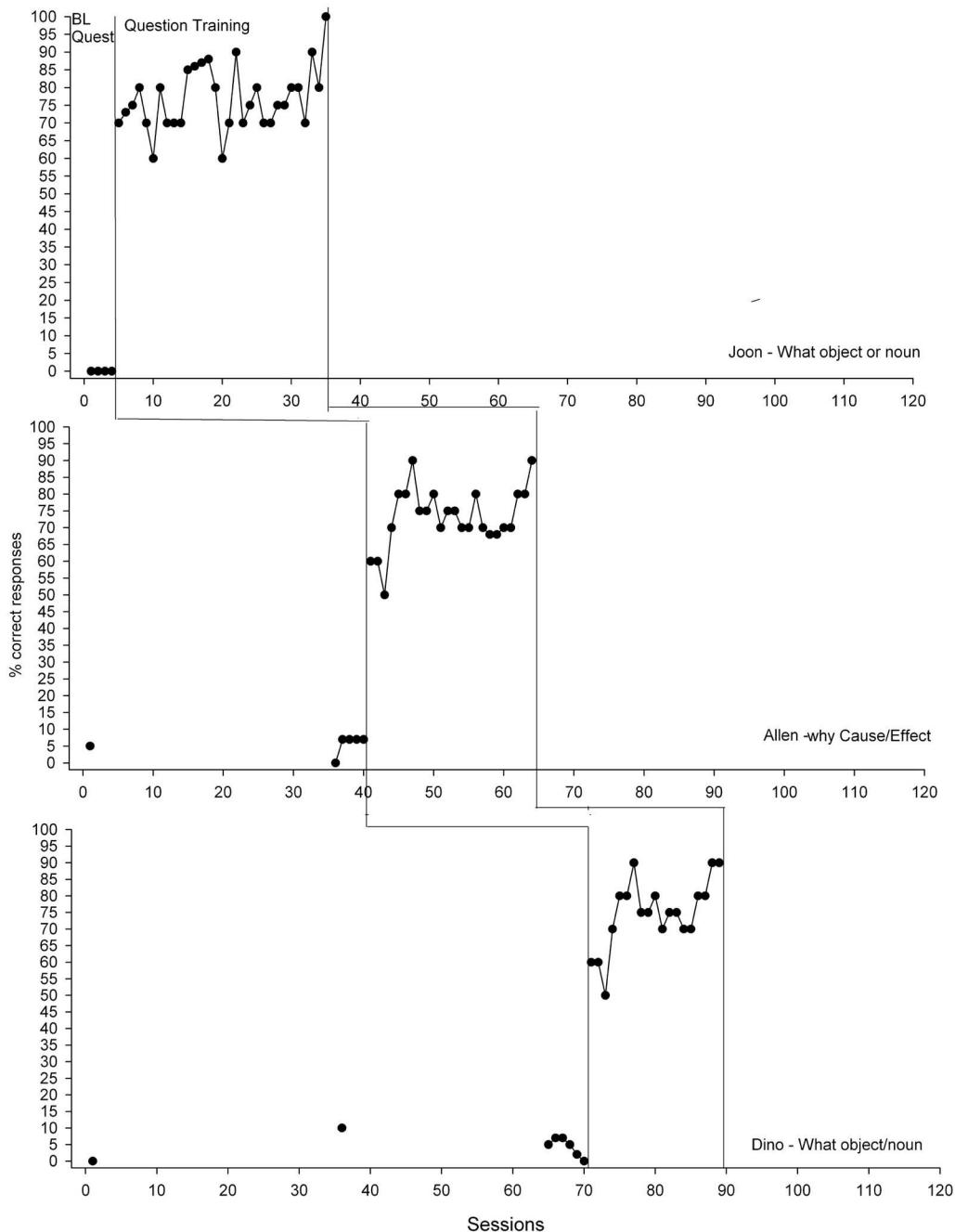


Figure 31.3 Sample Multiple Probe Graph.

on each remaining tier. Additional baseline probes are collected on tier two until there is at least one more probe session than occurred in tier one. Horner and Baer (1978) referred to these consecutive probes as the “true” baseline sessions and they always preceded the implementation of intervention. Thus, each tier has at least one more “true” baseline session than the preceding tier. An example of a multiple probe design is presented in Figure 31.3 and can be seen in the study by Lo, Correa, and

Anderson (2015). Lo et al. used a multiple probe design to assess the impact of culturally responsive social skills instruction on the social interactions of Latino and non-Latino students while at recess. The results of the study showed that each Latino student increased their frequency of appropriate social interactions with their non-Latino peers.

Multiple probe designs are more efficient than a multiple baseline design because baseline data are collected intermittently. However, this also presents a problem because a researcher would not see abrupt changes that might occur during baseline.

A second variation of the multiple baseline design is a *changing criterion design* (Hartman & Hall, 1976; Tawney & Gast, 1984). The design is typically used to evaluate behaviors that will increase in a gradual, stepwise fashion. Implementation of the design requires a baseline phase during which data are collected to show both the pre-intervention level of the behavior and to determine an initial criterion level (e.g., the average level of responding). Intervention is implemented subsequently until the target behavior reaches the first criterion level. Once the first criterion level has been reached, a more stringent criterion is established and intervention continues until that second criterion has been met. This pattern continues until responding is at the terminal criterion. The data collected at each criterion level serve as a baseline for the subsequent phase. Experimental control is demonstrated when behavior changes to the new criterion level each time the criterion is changed. The change should occur rapidly and responding should stabilize at the specific criterion level before the criterion is changed. An example of a changing criterion design is found in Figure 31.4. An example from the literature can be found in the study by Warnes and Allen (2005). Warnes and Allen assessed whether electromyographic (EMG) biofeedback could affect paradoxical vocal fold motion (PVFM) in a 16-year-old participant. Baseline levels of muscle tension were recorded initially followed by intervention with EMG biofeedback. Once lower muscle-tension had occurred and maintained at the criterion level, the criterion was changed and treatment continued until the new criterion level was met. This continued in a changing criterion design format until typical levels of muscle tension were attained. The results demonstrated that the intervention was effective in reducing muscle tension.

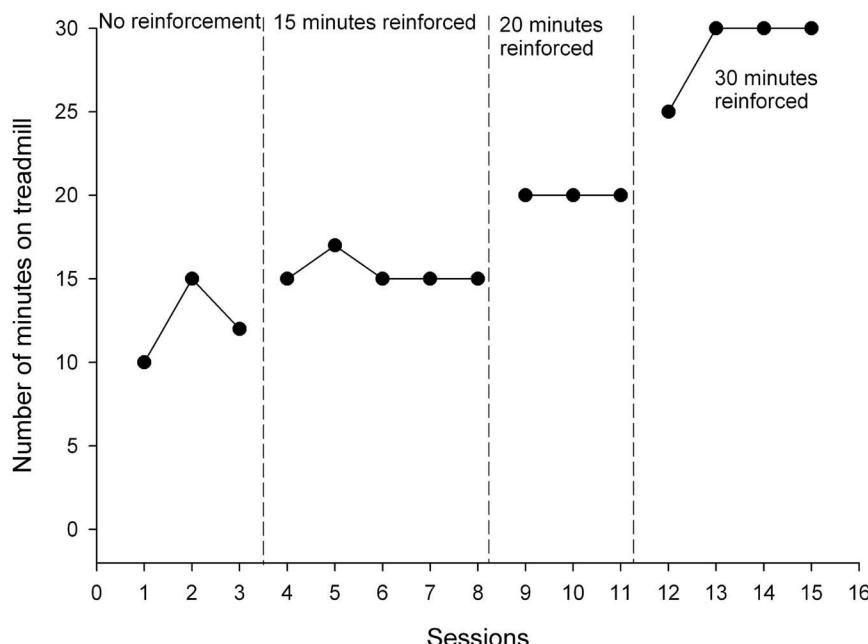


Figure 31.4 Sample Changing Criterion Design.

Alternating treatments designs (ATDS), also known as *multi-element designs*, *multiple schedule designs*, or *randomized designs*, are also considered to be an extension of the reversal design. In an alternating treatments design, two or more independent variables are alternated rapidly so as to compare their differential effects. A distinct stimulus is always associated with each condition to facilitate the participant's discrimination. This helps reduce the possibility of carryover effects and increases the likelihood that differences in responding are a function of different levels of effectiveness and not the participant's inability to discriminate the presence of the different conditions. Unlike the designs discussed above, an ATD does not require baseline data, a reversal, or stability in order to demonstrate experimental control. Fractionation, or salient changes in level between the two conditions, signifies effects (Barlow & Hayes, 1979; Tawney & Gast, 1984).

When alternating between conditions, it is best to randomize the presentation of the two conditions to avoid sequence effects. Ulman and Sulzer-Azaroff (1975) recommended that the presentation be random with no more than two of the same conditions presented in succession.

Other concerns for researchers include multiple treatment interference, generalization, and contrast effects. With generalization or carryover effects, responding is diffused among conditions by virtue of exposure. This is likely to occur when participants do not discriminate between the conditions. In contrast effects, the presence of one condition serves to suppress responding in another condition. Generalization and contrast effects can be avoided by making conditions salient. The researcher might assign certain uniforms or colors of stimuli to each condition. A classic article that utilizes an ATD is by Iwata, Dorsey, Slifer, Bauman, and Richman (1994). Iwata et al. sought to determine the relation between self-injurious behavior (SIB) and specific environmental events in an effort to identify the function of different forms of SIB prior to intervention. Participants were observed in four different conditions that were introduced in an ATD fashion: social disapproval, academic demand, unstructured play, and a final condition where the participants were placed in a room without access to any materials or persons. The results showed that, for the majority of participants, SIBs were consistently associated with one of the conditions listed above. These findings

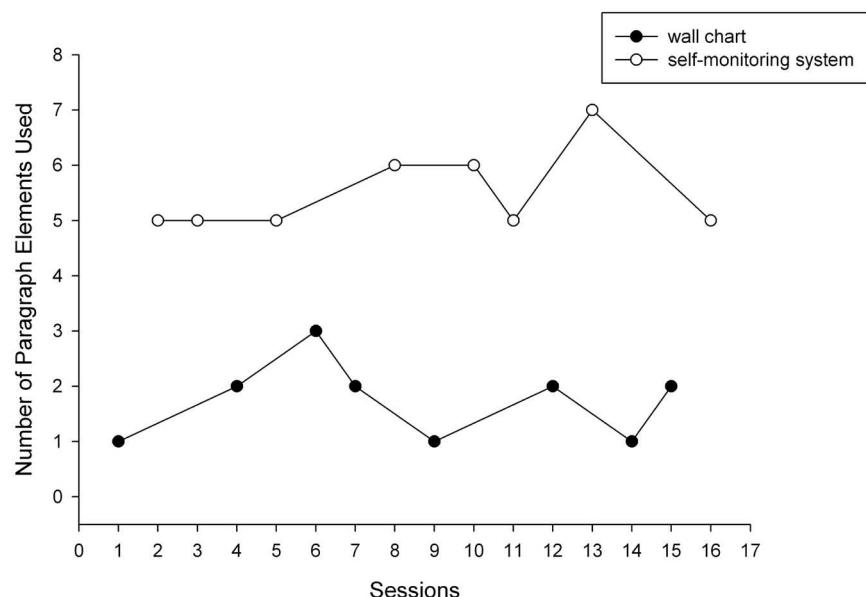


Figure 31.5 Sample Alternating Treatments Graph.

provided evidence that SIB may be a function of different sources of reinforcement. A sample alternating treatments graph is presented in Figure 30.5.

Variations on designs. In rare cases where studies produce novel results or are implemented in environments where strong environmental control is difficult to impossible, it is acceptable to vary from these basic designs as long as the most experimental control possible is achieved. An example of such a variation is non-concurrent multiple baseline design. The design differs from the multiple baseline design in that baselines are introduced one at a time (in an A-B fashion) and are not connected to the data collected in other tiers. Participants are randomly assigned to baseline conditions. For example, Harvey, May, and Kennedy (2004) suggested that a non-concurrent multiple baseline may be ideal for studying interventions applied to a whole school district. Taking baselines and implementing interventions in a sequential fashion across a large number of schools and locales could be unfeasible. Instead, districts could be added each year to show replication of effect (Harvey et al., 2004).

Choosing such a design variation will lead to less experimental control because nonconcurrent multiple baseline designs do not control for the effects of history and/or maturation (Christ, 2007; Harvey et al., 2004). This prevents a researcher from concluding that the independent variable per se was responsible for changes in the dependent variable. In these cases, the experimenters must explain clearly why they were unable to use a more controlled experimental design in the discussion section.

It is also common to see designs combined. For example, a researcher may elect to do a multiple baseline across subjects that includes a reversal. Combined designs allow for stronger conclusions about the effects of the intervention. Readers are encouraged to consult resources such as Kennedy (2005) for more information.

Experimental designs are selected based upon the salient aspects of the research question(s). In addition, dimensions of the operational definition as well as the constraints of the setting should be considered when selecting a particular research design. Ethical factors, such as the feasibility of withdrawing the independent variable, should be considered as well.

4. Replication

Replication is a critical feature of single-subject designs because it is through replication that both internal and external validity are demonstrated. Replication can take many forms. *Direct replication* is the replication of the procedures by the same experimenter using the same procedures either within or across participants. Sidman (1960) stated that there are two types of direct replication: *intrsubject*, where a participant is exposed to the intervention at multiple points in time (e.g., a reversal design), and *intersubject*, where a new participant is introduced to the study to replicate effects.

Systematic replication is a form of replication in which some part of the original study is altered to increase the generality of the results. For example, a conflict-resolution strategy that has been shown in the literature to be effective with adults might be applied to an adolescent population. Systematic replication typically occurs in a new study and with new researchers.

Replication in investigations using single-subject designs is also used to establish the external validity of experimental findings. Birnbrauer (1981) argued that the external validity of data from single-subject designs is established through systematic replication of effects by other researchers in settings different from the site in which the original study occurred. The focus in research using single-subject designs is to not only demonstrate the replicability of findings but to also establish the parameters of the intervention.

5. Independent and Dependent Variables

Researchers using single-subject research designs must implement the independent or treatment variable (IV) repeatedly over time. Ensuring that the IV is delivered reliably necessitates that it be

operationally defined, which requires researchers to describe all components in concrete, observable, and measurable terms. An operationally defined IV increases the likelihood that it will be implemented with consistency and, as a result, will strengthen the ability to demonstrate functional control. Moreover, it will aid other researchers in replicating the study procedures, which will be discussed in a later portion of this chapter.

The quality of an operational definition can be determined by collecting data on the implementation of the IV. Measuring fidelity of implementation (FOI) (i.e., treatment fidelity, treatment integrity) ensures that the IV is implemented accurately and consistently as it is defined or intended. According to Lane, Bocian, MacMillan, and Gresham (2004), measuring FOI may be done through: direct observation, feedback from consultants, self-monitoring and reporting, review of permanent products, and treatment manualization. High treatment integrity would provide evidence that the IV was well-defined. Gresham (2009), Hagermoser Sanetti and Kratochwill (2009), and Smith, Daunic, and Taylor (2007) are useful resources that discuss treatment fidelity. Ledford and Gast (2014) further identified specific procedures to measure procedural fidelity that emphasizes the collection of such data *across* conditions (e.g., baseline, intervention, post-intervention, comparison conditions) and not simply in the phase in which the intervention is being implemented.

Furthermore, it is critical that researchers using single-subject designs provide operational definitions for dependent variables (DV). Operationally defined DVs allows for accurate and consistent collection of data. The calculation of interobserver agreement (IOA) (i.e., interrater reliability) is discussed in detail later in this chapter. In the same manner that operational definitions of IVs facilitate replication, operational definitions of the DVs also facilitate future replication by other investigators, because what was measured in the original study is made transparent.

Data collection procedures. There are several methods that have been used by single-subject researchers to collect data. Selection of the method should be based upon the specific research question, the operational definition of the target behavior(s), and practical/environmental factors. If a data collection procedure is selected that does not reflect the operational definition and research question, is difficult to record (e.g., a child who taps a pen at such a high rate that counting each individual pen tap would be impossible), or there are restrictions on data collection (such as not having the ability to videotape a language sample), data taken might not be representative of the environment. Therefore, data collection procedures should be selected that maximize the probability that the data collected are representative of the actual environmental events.

The length of the observation will be dependent upon the probability that the behavior will be observed. If the response is occurring consistently throughout the day, then a shorter sample will more likely be representative of environmental events. Data collection can also be restricted to times when behavior is most likely to occur (e.g., if a child is most likely to have a tantrum after getting off of the bus, then data collection can be restricted to that particular time). If responding is more diffuse or less probable, the researcher will need to lengthen the data collection period to insure a representative sampling of the dependent variable.

One form of data collection used by single-subject researchers is *frequency*. Frequency data are collected on behaviors that have discrete onset and offset and occur at a distinct moment in time, in other words, behaviors that are “countable.” For example, the number of times a person checks email during office hours is an example of a behavior that would be measured using frequency data. Sometimes a researcher will know in advance that the target behavior occurs more/less frequently during a specific time period within an observation. In this situation, the time period would be broken into equal intervals, and the frequency of the behavior would be measured within each interval.

Duration data can be used to measure how long a behavior occurs. Tawney and Gast (1984) suggested that duration data could be collected in two ways depending on the characteristics of the target behavior. *Total duration* refers to the total amount of time a participant is engaged in a

defined behavior and can be used when the target behavior occurs continuously. For example, one might be interested in recording the total amount of time a student spent studying in an evening. *Duration per occurrence* is a measure of all contiguous occurrences of behavior. In the last example, a researcher would use duration per occurrence if studying was constantly interrupted, and s/he was interested in the duration of each discrete study period.

Latency data collection is used when an investigator is concerned with the length of time between the cue to respond and how long it takes for responding to begin, as opposed to how long it takes to complete a task. In the latter scenario, duration recording would be more appropriate. Latency recording would be appropriate, for example, if the researcher were interested in measuring the amount of time it took for a student to begin lining up at the door following a teacher's direction to engage in that behavior.

In some cases, the dependent variable might not have a clear onset or offset, or might occur too rapidly for reliable data collection using the methods mentioned previously. For example, if an experimenter is taking data on conversational skills with multiple components, it can be difficult to discern precisely where the interaction begins and where it ends. In this case, recording methods such as *partial interval*, *whole interval*, or *time sampling* might be appropriate. Interval measures require that observation periods are broken down into equal intervals (sometimes also known as *bins*). For partial and whole interval, time is split between observation intervals that are typically 10–20 seconds in length, and recording intervals that are typically 3–5 seconds in length. During observation intervals, the observer is monitoring for the occurrence of the dependent variable. No data are recorded until the end of the observation interval; conversely, no responses emitted during the recording interval are applied to the observation interval. If the observer is using a partial interval system, the dependent variable is recorded as occurring if it is emitted at any time during the observation interval. For whole interval data collection, responses are recorded *only* if the response is emitted for the entire observation interval.

Time sampling is used when continuous observation of the dependent variable is either impractical or unnecessary. In time sampling, time is also broken into equal intervals. Data are only recorded in the instant that the interval expires. The observer marks a "+" on the data sheet if the participant is engaged in the response at that moment. If the participant is not engaged, the observer marks a "-." Although some clinical and educational texts (e.g., Alberto & Troutman, 2013) state that time sampling intervals can be larger than partial or whole interval measures, for purposes of research it is important to note that intervals larger than 30 seconds may decrease the accuracy of the data. In cases where observations are 10 minutes or less, intervals should be about 10 seconds in length (Devine et al., 2011).

When using any of the above interval measures, the length of observation per session will depend upon the topography of the behavior. For example, high frequency behaviors of short duration would require shorter time intervals (10–20 seconds) while behaviors that are of low frequency with long durations would require longer observation intervals. It is important to note that larger intervals tend to be less sensitive to changes in responding. This may lead to an unacceptable inflation or deflation of the sample (Schmidt et al., 2013). Intervals larger than 2 minutes have been shown to decrease the validity of the data (Cooper, Heron, & Heward, 2007).

6. Inter-observer Agreement Coefficients

Researchers employing single-subject designs have historically used human observers to record the occurrence or nonoccurrence of behavior. Human observers, however, increase the likelihood of variability during observations (Kazdin, 1977). As a result, procedures were developed for measuring whether or not independent observers record data in a reliable manner (i.e., the operational

definition is interpreted consistently). Inter-observer agreement scores are typically reported as an overall percentage of agreement together with the range across experimental sessions. The literature recommends that inter-observer agreement be assessed on at least 30% of sessions, although this figure did not evolve from any research investigations. An acceptable level of observer agreement is usually 80% or above; however, this figure has also not been established through systematic research.

Different formulas have been developed for calculating inter-observer agreement depending on the type of data collected (see Chapter 10, this volume). Occasionally, an agreement coefficient such as *kappa* (Cohen, 1960) is calculated in single-subject research; however, the following formulae are seen most frequently in the single-subject literature. *Total agreement* (Kennedy, 2005) is typically used for *frequency*, *duration*, and *latency* data. The formula used to calculate total agreement between two observers is:

$$(\text{Smaller total of behavior recorded} / \text{Larger total of behavior recorded}) \times 100\%$$

One major limitation of this formula is that it does not take into account whether or not two observers ever agreed on the occurrence of individual instances of behavior. As a result, high levels of agreement between observers may occur, even though they have never agreed on the occurrence of a single behavior. Bailey and Burch (2002) suggested that calculating a block by block percent agreement and then averaging the scores would be one way to correct for the above problem.

Calculating percent agreement scores for data collected using *interval* measures (e.g., partial interval, whole interval, and time sample) involves a different formula that compares observer recordings interval by interval. The formula used to calculate percent agreement in this manner is:

$$(\text{Agreements} / (\text{Agreements} + \text{Disagreements})) \times 100\%.$$

An agreement was scored if both observers recorded the occurrence or nonoccurrence of a behavior in the same interval; disagreements occurred when one observer recorded an occurrence of a behavior and the second observer did not.

A more stringent method for determining agreement for interval type data is to use the same formula to calculate reliabilities for the occurrences and nonoccurrences separately. For example, percent agreement for occurrence data would be calculated using the formula:

$$[(\text{Agreements of Occurrence}) / (\text{Agreements of Occurrence} + \text{Disagreements})] \times 100\%$$

Percent agreement for nonoccurrence data would be calculated using the same formula except the focus would be on agreements of nonoccurrence.

7. Presentation and Summarization of Data

The visual analysis of graphic data both within and across conditions of a research study represents the most frequently used data analysis strategy employed by researchers using single-subject designs. Each single-subject design requires a specific pattern of data in order for an investigator to conclude that implementation of the independent variable is responsible for changes in the dependent variable. Researchers typically evaluate several factors when visually analyzing data: *Trend*, *level*, *variability*, *immediacy of effect*, *percentage of overlap*, and *consistency of data patterns across similar phases*. *Trend* refers to the slope of the data, and should not be confused with trend analysis typically used with inferential statistics. The trend of the data should be evaluated both within and between conditions.

A minimum of three data points must be collected in each condition in order to establish level and trend. This is especially true in baseline, which Birnbrauer (1981) equated to the descriptive data often seen in group designs. That is, baseline describes the participants' responding under naturalistic conditions before the application of the dependent variable. More data might be taken

in cases where level and/or trend are not easily evaluated (such as with highly variable data) or in the case where extended baselines show the independence of baselines, such as in a multiple baseline or multiple probe design. Fewer than three data points might be obtained in cases where extending a baseline or treatment condition might result in physical or psychological harm, such as in the case of severe self-injury.

Data should be stable (that is, three data points with little change in slope) or trending in the opposite direction to the desired effect before moving to another condition. Whether the desired trend is ascending or descending depends upon the conditions in place and the operational definition. For example, if a reversal design is used to determine the effects of praise on homework completion, one would want a stable or descending trend in baseline before proceeding. If homework completion was ascending during baseline and continued to increase during intervention, it would be difficult to determine whether the treatment per se was responsible for changes in responding. However, if the dependent variable was calling out, the investigator would want to have a stable or ascending baseline before proceeding.

The investigator should also see changes in trend between conditions as well. For example, few conclusions about the effectiveness of the dependent variable can be made if an ascending trend is observed in baseline and continues in treatment. However, if baseline responding is stable or descending and treatment results in an increasing trend, the independent variable clearly had an effect on the dependent variable.

Level refers to the height of the data on the *y* axis. Kennedy (2005, p. 197) referred to level as “the average of the data within a condition.” Level of data is often used to make a comparison between two phases. Immediate changes in level upon changes in condition suggest that the application of the treatment variable(s) is responsible for changes in the dependent variable.

Variability refers to the patterns observed between individual data points. Most data will not have a clean ascending or descending trend; however, the less variable the data, the more clear the pattern. Occasionally, data will be variable/stable, in which case a pattern of variability can be easily seen. For example, if two data points are ascending and one is descending, and this same pattern is observed at least once more, the data are considered to be *variable/stable*. It would be appropriate to move to the next phase of the investigation if this were the case. When data points are scattered in such a way that no pattern can be discerned, the researcher must attempt to control for the variability in the data. Unless the source of variability cannot be determined, the investigator should avoid moving to the next phase of the investigation.

Immediacy of the effect was defined by Kratochwill et al. (2010, p. 18) as “the change in level between the last three data points in one phase and the first three data points of the next.” A researcher’s conclusions about the effects of the independent variable can be much stronger if there is a rapid or immediate change in behavior once the intervention is applied.

Percentage of overlap is described by Kratochwill et al. (2010, p. 18) as “the proportion of data from one phase that overlaps with data from the previous phase.” It is critical to evaluate the overlap between baseline and intervention data because it could affect conclusions. The researcher can conclude that the intervention is effective if there is very little overlap between baseline and intervention data.

Consistency of data in similar phases refers to the similarity of data patterns across the same phases (Erchul & Sheridan, 2008). Evaluating consistency can be done by looking at data from all phases within the same condition, and then examining the extent to which there is consistency in the data patterns (trend). The greater the consistency, the more likely the data represent a causal relation.

It is crucial that changes in conditions (e.g., moving from baseline to treatment) are made based upon careful analysis of the trend, level, variability, immediacy of effect, percentage of overlap, and consistency of the data. Historically, inferential statistics have not been employed frequently to analyze data from single-subject designs; instead, visual inspection of data has been used. Functional

relationships are documented and analyzed by looking at patterns in the data. Statements such as “data were collected for three days for all participants” suggest that the schedule of the application of the independent variables were determined before the study began, and that changes in data patterns were not considered.

The description of the data in the narrative should match the graph, but be written with enough detail that it can stand by itself. Any changes in trend and level of the data should be described in detail. The mean and range of responding per condition is typically presented, and can help the researcher understand how responding changes over time.

8. Statistical Analyses of Data

As mentioned previously, visual analysis has predominantly been the method used to analyze graphed data from single-subject research. However, there has been a burgeoning emphasis on using statistical analyses. The two-sided debate surrounding whether statistical applications should be applied to single-subject research has been ongoing for the past 40 years (Kratochwill & Levin, 2014).

Researchers in support of the continued usage of visual analysis argue that the method allows for conservative judgements on intervention effects that take into account the nuances of single-subject research, especially in regards to level, trend, variability, and overlap of data patterns across phases. Additionally, Schull stated that “the fact that scientific decision making is often subtle, complex, and multifaceted, and faithfully following traditional statistical methods in a rule-governed fashion may limit or interfere with effective scientific judgments” (as cited in Fisher & Lerman, 2014, p. 245).

Furthermore, the use of statistics in single-subject research is in its infancy, and is still being constructed and tested. The field awaits further research to determine the most accurate and practical methods for detecting treatment effects that captures the complexities of this research methodology (Parker, Hagan-Burke, & Vannest, 2007). Additionally, there is little agreement on which statistical analyses should be applied to single-subject research. Horner et al. stated:

to date, however, no statistical approach for examining single-case research has met three fundamental criteria: (a) controls for auto-correlation (e.g. the fact that scores are not independent), (b) provides a metric that integrates the full constellation of variables used in visual analysis of single-case designs to assess the level of experimental control demonstrated by the data, and (c) produces an effect-size measure for the whole study (as opposed to two adjacent phases).

(Horner et al., 2012, p. 270)

Furthermore, Wolery (2013, p. 39) argued that researchers who vehemently apply typical statistical analysis “for data collected under different assumptions and to use indices familiar to them” are doing so through the violation of statistical assumptions.

However, methodologists such as Kratochwill and Levin (2014) suggested limited research substantiating the usage of visual analysis, and the reliability, validity, and accuracy of the method has generated the need for statistical analysis. Furthermore, Gast (2010) noted there are many advantages of adding statistical analysis to single-subject design including the detection of variables whose impact might be important but too small to be observed by visual analysis as well as the identification of intervention effects despite the presence of variability in the data. Additionally, Gast believed that statistical analysis can be more objective than visual analysis.

Utilization of statistical analyses with single-subject research may address the demand for summaries such as meta-analyses and research syntheses (Horner, Swaminathan, Sugai, & Smolkowski, 2012; Shadish, Rindskopf, & Hedges, 2008). Some researchers also feel that statistical analysis

increase the rigor of single-subject research; however, this notion is controversial and debated amongst researchers (e.g., Wolery, 2013).

Kratochwill et al. (2010) identified nonparametric methods—specifically, percentage of nonoverlapping data (PND), percentage of all nonoverlapping data (PAND), and percentage exceeding the median (PEM)—as being common methods used to systematically analyze or calculate effect sizes in single-subject research data. PND is determined by identifying the highest data point in baseline and calculating the percentage of data points which surpass this level during intervention (Scruggs, Mastropieri, & Casto, 1987). PAND is calculated by identifying the total number of overlapping points and dividing it by the total number of points and subtracting the percentage from 100 (Parker et al., 2007). PEM is determined by identifying the median data point in baseline and calculating the percentage of data points above this level, if the dependent variable data is expected to increase, and below this level if the dependent variable data is expected to decrease (Ma, 2006).

Other nonparametric methods (e.g., pairwise data overlap, nonoverlap of all pairs, improvement rate difference, percentage of data exceeding a median trend, tau-U, extended celeration line), regression estimates (e.g., 4-parameter model) and multilevel models (e.g., hierarchical linear models) have also been used to statistically analyze single-subject data (Hott, Limberg, Ohrt, & Schmit, 2015; Kratochwill et al., 2010; Parker, Vannest, & Davis, 2011; Rakap, 2015; Vannest & Ninci, 2015; Wendt, 2009). However, it is important to reiterate that the above mentioned statistical tests have both strengths and limitations that may influence the drawing of conclusions.

Single-subject designs are useful for research where individual variability is the unit of analysis. Good studies should include designs that sufficiently isolate the independent variable and include repeated measures of the dependent variable, measures of treatment fidelity, and consistent interpretation of the independent variable. Replication across time, settings, individuals, and/or materials is critical for showing effects as well as demonstrating the generalizability of the findings. The type(s) of design, data collection, and analyses chosen will depend upon the research question and the variables studied.

Acknowledgment

Preparation of this chapter was supported in part by USDE OSEP grant #H355A040025.

Note

- 1 An A-B-A is the basic form of the reversal design. Although experimental control can be established using this form of the design, researchers prefer to use the A-B-A-B because it allows for additional replication and ends on an intervention phase.

References

- Ahearn, W. H., Clark, K. M., MacDonald, R. P. F., & Chung, B. I. (2007). Assessing and treating vocal stereotypy in children with autism. *Journal of Applied Behavior Analysis*, 40, 263–275.
- Alberto, P. A., & Troutman, A. C. (2013). *Applied behavior analysis for teachers* (9th ed.). Upper Saddle River, NJ: Merrill-Prentice-Hall.
- Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, 1, 91–97.
- Bailey, J. S., & Burch, M. R. (2002). *Research methods in applied behavior analysis*. Thousand Oaks, CA: Sage.
- Barlow, D. H., & Hayes, S. C. (1979). Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis*, 12, 199–210.
- Birnbrauer, J. S. (1981). External validity and experimental investigation of individual behavior. *Analysis and Intervention in Developmental Disabilities*, 1, 117–132.
- Briere, D. E., Simonsen, B., Sugai, G., & Myers, D. (2015). Increasing new teachers' specific praise rates using a within-school consultation intervention. *Journal of Positive Behavior Interventions*, 17, 50–60.

- Christ, T. J. (2007). Experimental control and threats to internal validity of concurrent and nonconcurrent multiple baseline designs. *Psychology in the Schools*, 44, 451–459.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Pearson Merrill Prentice Hall.
- Devine, S. L., Rapp, J. T., Testa, J. R., Henrickson, M. L., & Schnerch, G. (2011). Detecting changes in simulated events using partial-interval recording and momentary time sampling III: Evaluating sensitivity as a function of session length. *Behavioral Interventions*, 26, 103–124.
- Erchul, W. P., & Sheridan, S. M. (2008). *Handbook of research in school consultation*. New York: Lawrence Erlbaum Associates.
- Fisher, W. W., & Lerman, D. C. (2014). It has been said that, “There are three degrees of falsehoods: lies, damn lies, and statistics.” *Journal of School Psychology*, 52, 243–248.
- Gast, D. L. (2010). *Single subject research methodology in behavioral sciences*. New York: Routledge.
- Gast, D. L., & Ledford, J. R. (2014). *Single case research methodology: Applications in special education and behavioral sciences* (2nd ed.). London: Routledge.
- Gresham, F. M. (2009). Evolution of the treatment integrity concept: Current status and future directions. *School Psychology Review*, 38, 533–540.
- Hagermoser Sanetti, L. M., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: Introduction to the special series. *School Psychology Review*, 38, 445–459.
- Hartman, D. P., & Hall, R. V. (1976). The changing criterion design. *Journal of Applied Behavior Analysis*, 9, 527–532.
- Harvey, M. T., May, M. E., & Kennedy, C. H. (2004). Nonconcurrent multiple baseline designs and the evaluation of educational systems. *Journal of Behavioral Education*, 13, 267–276.
- Horner, R. D., & Baer, D. M. (1978). Multiple-probe technique: A variation on the multiple baseline. *Journal of Applied Behavior Analysis*, 11, 189–196.
- Horner, R. H., Carr, E. G., Halle, J. W., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, 71, 165–180.
- Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children*, 35, 269–290.
- Hott, B. L., Limberg, D., Ohrt, J. H., & Schmit, M. K. (2015). Reporting results of single-case studies. *Journal of Counseling & Development*, 93, 412–417.
- Iwata, B. A., Dorsey, M. F., Slifer, K. J., Bauman, K. E., & Richman, G. S. (1994). Toward a functional analysis of self-injury. *Journal of Applied Behavior Analysis*, 27, 197–209.
- Jones, K. M., & Friman, P. C. (1999). A case study of behavioral assessment and treatment of insect phobia. *Journal of Applied Behavior Analysis*, 32, 95–98.
- Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis*, 10, 141–150.
- Kazdin, A. E. (1981). External validity and single-case experimentation: Issues and limitations (A response to J. S. Birnbrauer). *Analysis and Intervention in Developmental Disabilities*, 1, 133–143.
- Kennedy, C. H. (2005). *Single case designs for educational research*. New York: Allyn & Bacon.
- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single-case designs technical documentation. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.
- Kratochwill, T. R., & Levin, J. R. (2014). *Single-case intervention research: Methodological and statistical advances*. Washington, DC: American Psychological Association.
- Lane, K. L., Bocian, K. M., MacMillan, D. L., & Gresham F. M. (2004). Treatment integrity: An essential but often forgotten component of school based interventions. *Preventing School Failure*, 48, 36–43.
- Ledford, J. R., & Gast, D. L. (2014). Measuring procedural fidelity in behavioural research. *Neuropsychological Rehabilitation*, 24, 332–348.
- Lo, Y., Correa, V. I., & Anderson, A. L. (2015). Culturally responsive social skill instruction for Latino male students. *Journal of Positive Behavior Interventions*, 17, 15–27.
- Ma, H. (2006). An alternative method for quantitative synthesis of single-subject researchers: Percentage of data points exceeding the median. *Behavior Modification*, 30(5), 598–617.
- Mayfield, K. H., & Vollmer, T. R. (2007). Teaching math skills to at-risk students using home-based peer tutoring. *Journal of Applied Behavior Analysis*, 40, 223–237.
- Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *The Journal of Special Education*, 40, 194–204.
- Parker R. I., Vannest K. J., & Davis J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification*, 35, 303–322.
- Rakap, S. (2015). Effect sizes as result interpretation aids in single-subject experimental research: Description and application of four nonoverlap methods. *British Journal of Special Education*, 42, 11–33.
- Saville, B. K., Zinn, T. E., Neef, N. A., Van Norman, R., & Ferreri, S. J. (2006). A comparison of interteaching and lecture in the college classroom. *Journal of Applied Behavior Analysis*, 39, 49–61.
- Schmidt, M. G., Rapp, J. T., Novotny, M. A., & Lood, E. A. (2013). Detecting changes in non-simulated events using partial interval recording and momentary time sampling: Evaluating false positive, false negatives, and trending: Detecting changes in non-simulated events. *Behavioral Interventions*, 28, 58–81.
- Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single subject research: Methodology and validation. *Remedial and Special Education*, 8, 24–33.

- Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, 3, 188–196.
- Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. Boston, MA: Authors Cooperative.
- Skinner, B. F. (1938). *The behavior of organisms*. New York: Appleton-Century-Crofts.
- Smith, S. W., Daunic, A. P., & Taylor, G. G. (2007). Treatment fidelity in applied education research: Expanding the adoption and application of measures to ensure evidence-based practice. *Education and Treatment of Children*, 30, 121–134.
- Tawney, J. W., & Gast, D. L. (1984). *Single subject research in special education*. Columbus, OH: Charles E. Merrill Publishing Company.
- Ulman, J. D., & Sulzer-Azaroff, B. (1975). Multielement baseline design in educational research. In E. Ramp & G. Semb (Eds.), *Behavior analysis* (pp. 371–391). Englewood Cliffs, NJ: Prentice-Hall.
- Vannest, K. J., & Ninci, J. (2015). Evaluating intervention effects in single-case research designs. *Journal of Counseling & Development*, 93, 403–411.
- Warnes, E. D., & Allen, K. D. (2005). Biofeedback treatment of paradoxical vocal fold dysfunction and respiratory distress in an adolescent female. *Journal of Applied Behavior Analysis*, 38, 529–532.
- Wendt, O. (2009). Calculating effect sizes for single-subject experimental designs: An overview and comparison. *PowerPoint slides*. Retrieved from www.campbellcollaboration.org/artman2/uploads/1/Wendt_calculating_effect_sizes.pdf
- Wolery, M. W. (2013). A commentary: Single-case design technical document of the What Works Clearinghouse. *Remedial and Special Education*, 43, 39–43.

32

Social Network Analysis

Tracy Sweet

Social networks are defined by a collection of individuals or entities (called nodes) and the relationships among them (called ties or edges). Examples of social network data include friendship ties among a group of people, collaboration ties among workplace employees, and co-authorship among a group of researchers. *Social network analysis* is a broad term used for any quantitative analysis of social network data. Such methods may be exploratory or descriptive in nature involving summary statistics or they may be inferential and involve some type of statistical model.

Social network data are often collected by survey. Respondents are asked to nominate other people about their interactions or relationships. For example, to generate a friendship network, each person may be asked to list their friends in their class. Network data can also be collected in other ways; respondents may be asked to construct the network, such as naming pairs or groups of individuals who are friends.

Social network data are often analyzed as binary because the methods for binary network analysis are more popular; however, valued ties are not uncommon. That is, rather than a dichotomous variable for whether two people are friends, an ordinal or continuous variable is used to measure the closeness of the relationship or frequency of interaction. Furthermore, social network ties may be *directed* in that a tie from node A to node B does not imply a tie from node B to node A. For example, teacher A seeks advice from teacher B, but teacher B may not seek advice from teacher A. Other ties, such as collaboration or co-authorship, imply a tie in both directions and these ties are called *undirected*.

Visual representations of network data include a plot called a *sociogram* in which each node is represented by a vertex and the ties as an edge between two nodes. Binary sociograms are common figures where the presence of an edge between two vertices corresponds to the presence of the relationship between two nodes (Figure 32.1, top). Another common representation for smaller networks is a plot of the $n \times n$ adjacency matrix, which is a matrix such that the entry in the i th row and j th column corresponds to the value of the tie from node i to node j . A visual representation of this matrix is a grid in which blocks are shaded according to the value of the entries in the matrix (see Figure 32.1, bottom).

Nodes and networks are commonly summarized by network statistics. Common node statistics include a variety of measures of centrality and include the number of ties going out of or in to

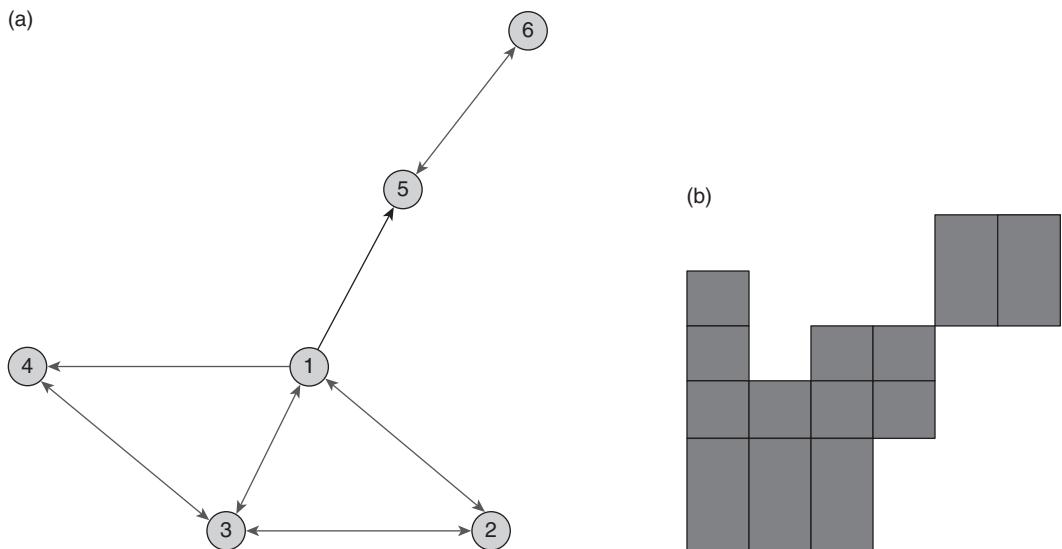


Figure 32.1 Visual Representations of Social Network Data: (a) a Sociogram and (b) Adjacency Matrix.

each node (called *out-degree* and *in-degree*, respectively). Other measures of centrality are based on path length, the number of edges needed to go from one node to another. *Betweenness* measures the extent to which nodes act as brokers; thus, nodes with high betweenness fall on path lengths between many nodes. *Closeness* is another node-level statistic that is a function of path length so that nodes with high closeness have short path lengths to other nodes.

Networks are also summarized by a variety of statistics as well as aggregates of node-level statistics. For example, the proportion of observed ties out of all possible ties is called *density*. Networks that have large proportions of ties are called *dense* networks and networks with relatively few ties are called *sparse* networks. Another measure, *reciprocity*, is determined by the proportion of directed ties that are reciprocated, that is, a tie from i to j appears with a tie from j to i . Other network statistics include the counts of various network structures such as triangles or k -stars (nodes with in- or out-degree of k). For a comprehensive list of other node and network statistics, see Wasserman and Faust (1994).

Because methods vary widely by discipline and because inferential network analysis is a relatively young field, there is not a standard line of analysis used in the social sciences. In some fields, it is quite common only to look at descriptive statistics or to do hypothesis tests based on the relative frequencies of certain network structures. In other fields, it is standard to fit quite complex statistical social network models.

The purpose of this chapter is to illustrate how to critically read social network research without an optimal method of analysis. Potential reviewers and researchers are therefore directed to common areas for possible criticism as well as questions that ought to be addressed by authors for any type of analysis. In addition, we include some introductory material about social network analysis as well as types of network models to aid with notation and terminology.

1. Social Network Theory

Because there is a myriad of methodological approaches to social network data, substantive theory should drive much of the research. For example, given any group of individuals, any number of

Table 32.1 Desiderata for Social Network Models.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. Theories (e.g., social, organizational) are presented to justify analyzing the particular type of network tie (e.g., friendship, advice-seeking, directed/undirected), as well as the type of analysis/model used.	I, M
2. The network data collection process is described in detail or the authors provide a reference to where that information can be accessed.	M
3. How the network data were cleaned and/or dichotomized is explained.	M
4. Missing network data are described and the treatment of missing data is explained.	M, R
5. If the network is modeled as an outcome, an appropriate social network model is used or another model that sufficiently defines the tie dependence structure is used. If the network is modeled as a predictor, the appropriate network or node summary statistic(s) are used and justified.	M, R
6. Model selection and model assessment, possibly by comparing competing models, are discussed.	M, R
7. Network data are summarized using standard network summary statistics and presented clearly in figures and/or tables.	R
8. If there are multiple networks, multiple types of ties, and/or multiple groups of nodes, they are explained explicitly and possibly displayed visually.	R
9. Relevant parameter estimates are provided with the appropriate intervals or error estimates; other variables in the model need not be included (e.g., latent space positions, block membership)	R
10. Correct inference is made with respect to the fitted model(s) and maps back to the original theory or research question(s).	R, D
11. Limitations of the models and analysis are discussed.	R, D

* I = Introduction, M = Methods, R = Results, D = Discussion.

possible relationships may exist, and different types of network connections yield both different types of network structures and support different inferential conclusions. Therefore, the type of network tie analyzed should be clearly motivated by theory. For example, friendship ties are a common form of network tie, but the research under review should offer some kind of social theory that explains why friendship is the optimal type of relationship for the research question.

Furthermore, how that tie is measured may also be motivated by theory. Although binary ties (relationship is present or absent) are most common, they provide the least amount of information about the value of the relationship. In many cases, the research question may require valued ties. For example, we may not only care whether two individuals are friends, but we also care about the level of closeness or frequency of interaction between those individuals. If the research question strongly suggests that valued ties are vital, it may be unreasonable to analyze data using binary ties. This suggestion is more obvious for exploratory analyses or if the network is then summarized as a predictor in another analysis as opposed to analyses where the network is being predicted through a network model which may not always accommodate valued ties (see Social Network Models desideratum for more information).

In addition to the type and value of the network tie, whether that tie is considered directed or undirected may also be a decision requiring explanation/justification. In some instances, the type of tie automatically implies that the tie is undirected; examples include co-authorship, collaboration, and physical proximity. In other contexts, whether the tie is considered by the

researchers to be directed or undirected may require further justification. For example, friendship is not necessarily an undirected relationship. Student A nominates Student B as a friend, but Student B does not nominate Student A. Researchers may choose to model friendship as a directed tie or they may decide only to count as ties those in which both students reciprocally nominated each other and analyze these ties as undirected. Therefore, explanation about how data are collected and how the network is constructed is important. For example, if the authors conduct an analysis in which only reciprocated ties count as friendship ties, the subsequent inference may be quite different from other studies in the field in which friendship ties are considered directed.

Because there may be a variety of ways to construct a network, we suggest that authors use social theory or even previous research to justify how they determine a tie. Similarly, theory should drive the analysis plan; which descriptive statistics or models will be used or which covariates will be explored should in large part be driven by the substantive theory described in the introduction. Similarly, how the network will be incorporated into a statistical model (if at all) is often determined by the substantive theory as well as the research question.

2. Data Collection Process

Social network data are generally collected via survey and the data collection process should be detailed because survey items for network data can vary widely. If the data collection process is not described, authors should cite the source that does include such a description.

There are several ways that surveys can solicit network information. Both the manner in which people are primed as well as whether names are automatically populated may affect the resulting network. The study should include the actual network survey item(s) and describe whether there was a previous name generation question. Name generation questions are common in social network data collection and descriptions of these items should be included. For example, survey respondents might first be asked to nominate individuals with whom they interact (e.g., “With whom do you interact on a regular basis?”) and this list is then used to populate future social network survey items (e.g., “To whom do you go for advice around science instruction?”). In addition, authors should include information about how names were populated and if not automatically populated, what information was solicited. For example, how many people did the item ask respondents to nominate, that is, how many blanks were there? Similarly, what if a respondent did not want to nominate anyone—how would that response be different from a missing response?

If the data collection process differs from other studies in the field, then the resulting network will likely differ from those studies. Generally, differences in surveying result in differences in the average number of people nominated, and this can result in the over- or under- appearance of certain network structures. For example, a survey with 5 blanks for friends will likely have data in which each node nominates no more than 5 other individuals, so the average number of out-ties will be less than 5. Another survey with 12 blanks for friends will likely result in data with average numbers of out-ties much larger than 5. These two networks will also differ in the number of other sub-structures as well, such as triads or stars. Similarly, the resulting network from 12 blanks will include ties that are likely much weaker than the 5 blanks. This not only affects the relationship conveyed by the network, it also will affect subsequent analyses because the networks themselves are fundamentally different.

The goal is for authors to be explicit about how network data are collected so that studies are comparable. If network data were acquired data in a less conventional manner, the results will likely differ from past research and it may not be possible to know if these differences are a result of the

study/individuals or simply a difference in the collection process. For studies involving multiple networks, the stakes are even higher because these studies aim to be more generalizable. If the data collection process varies across sites, these networks are no longer comparable and this has implications for analysis.

3. Data Processing

Network data are commonly displayed and analyzed as binary because many network statistics and models primarily use binary network data; however, depending on the survey item, the network ties may actually be ordinal or continuous. How these data are then dichotomized should be explained given that different methods of dichotomization result in different types of networks (and potentially different inference). Authors should justify their decisions, perhaps through some kind of data exploration or based on theory. For example, teacher advice-seeking ties may be ordinal and include a frequency measure, but for a write-up authors may decide to include advice-seeking that occurred at least once each month as a tie based on an assumption that social capital required sustained relationships for the exchange of social resources.

Similarly, if the network data are not the direct result of nominations, the research should describe how the data were processed to construct a network. For example, researchers may decide that a friendship network tie requires mutual nomination; that is, if node A nominates node B as a friend but B does not nominate A, then a friendship tie does not exist between A and B. This decision process should be explicit to readers. Another example is if some studies ask all respondents to list the ties among all individuals in the network, which results in multiple nominations of ties. Then authors should describe how many nominations (or preferably what proportion of nominations) result in a tie. These decisions affect the ultimate network structure and the general type of relationships represented by the resulting network.

If future analyses will use different forms of the network data, such as a network plot will use dichotomized data but measures of centrality will use continuous tie data, authors must be extremely careful in communicating which network data are used where and should include some form of justification for introducing additional complications into the analysis.

It is also possible that authors do not have credible justification for their decisions. If the work is more exploratory in nature, lack of justification does not necessarily imply a weakness of the research, but a reviewer may ask authors to construct their network in another way for comparison. Otherwise, the results and conclusions should be taken with caution.

4. Missing Data

Missing data is an issue in all data analysis, but the dyadic nature of networks complicates how missing data is handled. One particular issue with missing social network data is that it is difficult to impute missing data because of the complex dependence structure among ties. In fact, most analyses of network data do not attempt to impute missing data, but missing data should be discussed.

In addition, social network missing data are unique because missing data for a given individual is usually only half missing. Missing data commonly results from not completing or taking surveys so that tie nominations from those individuals are missing. Someone failing to take a survey however does not preclude someone else from nominating them. Thus, given a missing survey from node A, we have information about ties being sent to A but not ties sent from A.

One option is to analyze the networks without ties coming from A but keeping A in the network. Thus, we then assume that A does not send any ties. In a network that is particularly sparse, including one or two of these missing nodes in the network would be acceptable;

in general, however, if there is a non-trivial proportion of nodes missing, they should not be included in the network.

Undirected networks have an advantage with missing data in that ties can be inferred. If the tie is undirected, such as friendship and collaboration often are measured, then network data around node A is ultimately determined from rest of the network. In a network with very few missing nodes, handling missing data in this way is generally acceptable.

Larger complications arise for both directed and undirected networks when large numbers of individuals are missing because the ties among these individuals are also missing. Including these individuals in the network and assuming none of them share ties is likely to impact results.

A more common approach is to simply exclude those individuals who did not respond to the survey from the network entirely. Depending on who and what proportion are missing from the network will determine what conclusions the authors can make. For example, if the authors can argue that the missing individuals are likely similar to the individuals who did complete the survey, inferences made should be the same had the entire network been available. Otherwise, authors should approach their conclusions more tentatively.

If authors do attempt to impute missing data, the imputation process should be described in detail. It is possible to impute ties either by imputing each missing tie with probability equal to the network density, but often this obstructs network structure. If manuscript authors impute missing ties using a social network model, the covariates included in such a model should be standard covariates used in other network models in that literature. In cases where network models are not common in the literature, imputation via network model is not recommended.

5. Network Models

The term *social network model* generally refers to modeling the network as the outcome so that covariates included in the model predict the presence or value of ties. In this section, we will also discuss models that include features of social networks as predictors but they are often considered a separate from other social network models.

Modeling social networks requires specialized models because networks ties are not independent of each other. Moreover, there is not a standard dependence structure that explicitly defines this dependence among ties. For example, friendship network ties are not independent of each other. Consider a friendship tie between A and B and between B and C. The probability of a tie between A and C should be greater (than a tie between any random pair of nodes) because A and C already share a friend. Similarly, if we know that teacher A seeks advice from B and C, then A may be less likely to seek advice from D given that teachers tend to have only a few advisors. Identifying which ties are likely to be dependent is a difficult process because of the community structure of networks.

Although the dependence structure is difficult to describe explicitly, it is obvious that these ties are not independent, so models that assume independent observations (e.g. logistic regression, classification methods) should not be used. Violations of this assumption can result in incorrect inference. Therefore, if the network is modeled as an outcome, an appropriate social network model should be used or another model that sufficiently defines the tie dependence structure is used. If the network is modeled as a predictor, the appropriate network or node summary statistic(s) is used and justified. There are several types of social network models, all of which can be augmented in a number of different ways. Thus, the reviewer's task is not to select which class of model ought to be used but to assess whether the model proposed is appropriate. Does the model accommodate a tie dependence structure? What are the assumptions of this model and are they appropriate for these data? These are questions that authors should be addressing in their manuscripts. For example,

dependence is defined explicitly in an exponential random graph model, also known as a p^* model, through choice of network structures included in the model. Authors would therefore need to justify their choice of network structures and explain how each structure accounts for tie dependence.

Other models, such as *latent space models*, assume ties are independent conditional on covariates and latent variables in the model. These models do not require justification about the dependence structure since these models were created for that purpose, but authors will still need to explain other aspects of their choice of model along with the relevant covariates. In addition, there are some models that focus on cluster or community structure such as *blockmodels* or *cluster extensions* of latent space models, and authors should convince the reader that these models are appropriate for their network data. For example, is cluster structure evident in the network visually or through some community segregation measure?

For some research questions, researchers may instead want to model the network as a predictor and estimate the impact (or influence) of the network on an individual (node-level) outcome. In this case, special attention must be paid as to how the network is embedded into a statistical model and what kind of inference is made. One major issue is that influence is generally confounded with homophily; that is, individuals with ties are likely to become more similar, but individuals who are similar are more likely to have ties. With cross-sectional data, we cannot discern the two. In many studies, it is impossible to disentangle these two phenomena, so authors must carefully explain how they are modeling influence.

Therefore, if authors are claiming that the network ties are associated with a change in some outcome, the authors should justify that the converse did not occur. For example, suppose a manuscript wants to relate friendship ties with smoking. If they survey students about smoking status and network ties at the same time, we don't know if students are friends with each other because they both smoke, or if students start smoking because their friends start smoking.

In most cases, homophily and influence can be disentangled if students are surveyed at different times in the year and longitudinal data are often preferred for these analyses. For example, if we have a smoking status collected at time 1, smoking status collected at time 2, and the existing friendship ties between times 1 and 2, we can fit a network model to look at the effect of smoking status on network ties to show that the homophily effect is null, that is whether both individuals are smokers or non-smokers does not predict ties. If we then look at the effects of the network at time 1 on smoking status on time 2, we can then assume that the effects of the network actually influence smoking as opposed to homophily. Without the two time points, however, we would not be able to draw such a conclusion. Note that there are more sophisticated longitudinal models aimed at disentangling homophily and influence (e.g., SIENA; Snijders, Steglich, Schweinberger & Huisman, 2008), but these models are quite complex and are not commonly used in many disciplines (e.g., education).

Assuming that authors are able to separate influence and homophily, common influence models in education research involve creating a variable that summarizes the network for each node. These node-level statistics are then used as covariates in some other statistical model, such as a linear model. The effect of the network is then estimated by the regression coefficient. Because of the sheer number of summary statistics that one can use, authors should explain in detail how they selected the summary statistic used in their model and a reviewer might ask authors to discuss (as a potential limitation) how the inference would have changed if another network statistic had been used. For example, one might be interested in the effects of degree on some node-level outcome variable. Degree, the number of ties connected to each node, is a measure of centrality, but authors should justify why they are using degree as opposed to in-degree, out-degree, betweenness or closeness for example.

In general, multiple node-level statistics are not used because network statistics are often highly correlated. If authors choose to include multiple network statistics in a linear model, reviewers should request additional diagnostic measures to ensure these covariates are not correlated. Further,

authors must justify their choice in including multiple summary statistics and show that models including these multiple measures are better in terms of model fit.

6. Model Selection and Model Assessment

If a social network model or other type of statistical model is used, some type of model selection process or model assessment procedure should be included in the manuscript. That said, methods for goodness of fit and model selection for social network models are active areas of research so the types of diagnostics used may appear elementary compared to other types of statistical models.

We propose that authors lean heavily on the literature to specify their models and use diagnostics and model comparisons as a way to amass evidence of an extremely poorly fitting model rather than as evidence to select the optimal model. For example, substantive theory should drive which covariates are included in the model. Depending on the social network model used, network structures may also be included in social network model and these should also be selected based on the assumed dependence structure of the network ties. In both cases, the reviewer should not expect a lengthy data driven model selection process that often occurs with regression models. Model fit indices such as AIC, BIC, and DIC are sometimes used but some research suggests that a model with optimal information criterion may in fact be a poor model (Handcock, 2003). Standard methods of goodness of fit include additional simulations to determine the likelihood of the observed network given the distribution of networks generated from the fitted model (Hunter, Goodreau, & Handcock, 2008), but for most substantive papers, including such methods is actually quite rare.

One alternative is to take advantage of the diagnostics particular to each model when possible. For example, latent space positions can be used as a diagnostic for latent space models to determine if the covariates included in the model capture enough of the network structure. If enough covariates are included in a latent space model, the latent positions function in a similar way as residuals in regression. Similarly, any type of clustering model could be compared with a clustering algorithm to assess the classification, but these additional analyses are not necessary if the authors have justified via prior studies which covariates/network structures should be included.

Another alternative method for model assessment is to compare parameter estimates to other studies in the literature or if prior research is unavailable, then to fit two different network models and compare parameter estimates. The second model would then serve as a diagnostic for the first.

We do note however that assessing model fit and assessing model convergence are two different procedures and that authors should include evidence of model convergence if fitting a model using an iterative procedure. Standard statistical software packages usually include constraints or checks to ensure convergence, but network models are often fit using open source packages and it falls to the analyst to ensure model convergence.

The consequences of not performing any type of model assessment are obvious because inference depends directly on the parameters estimated from these models. There should be some evidence that parameter estimates are sound (previous research or model comparison). Similarly, if model estimation algorithms are not converging, the parameter estimates could be incorrect.

7. Descriptive Summaries and Visualization

Summaries of common network statistics should appear as part of the Results section in which the data are described and these statistics should appear in tables and/or with figures. Summary statistics such as network size and density are most common but other statistics may be necessary based on the analysis.

Summaries of networks not only provide readers with insight about network structure but also motivate subsequent analyses. For example, showing that the number of triangles (three connected nodes) is particularly high in one network may be relevant in explaining why that same network has low overall density and could be justification for applying a different model or analysis to this network.

Network data are often represented visually using a sociogram in which vertices and edges represent nodes and ties respectively. These figures should be labeled appropriately and described in full detail. Because tie values can be binary, ordinal or continuous, any variation in edge-width should be described. If the manuscript is submitted to a journal that does not regularly publish papers with social network data, then it is also necessary that authors describe the relative positioning of the nodes on the sociogram. Note that most network software will employ some kind of default algorithm to space the nodes but because default settings vary, authors should specify either the algorithm the network was plotted or explain that the relative distance between nodes is generally meaningless. If the latter is not true, the authors should state this explicitly.

Visual representations are often necessary as they also motivate the type of analysis or model being used. For example if a network displays subgroup structure, subsequent analysis should accommodate such structure at the very least if not be fundamentally guided by it. If the network is too large, often subsets of the network are shown visually.

8. Multiple Networks

Often in the social sciences and in the education sciences in particular, data include multiple networks such that each network contains different individuals, such as teachers in different schools. Another type of multiple network includes different types of relationships among the same group of individuals, and a third type involves the same type of relationship among the same group of individuals but at different times. All of these examples are called *multilevel networks* in different fields. Because the literature is not consistent with nomenclature, the authors must be clear about the type of network data collected.

In the case of multiple independent networks, such as teachers in separate schools, networks should be summarized and displayed separately to illustrate the fact that these networks are isolated from one another. Typically a space, line or box is used to separate different sociograms. Because these networks are independent, the authors should conduct separate analyses to compute summary statistics rather than aggregate the nodes and ties. If authors attempt to analyze these networks as a single network, the network will be overly sparse since ties across schools are rarely collected. There are exceptions in which multiple schools are analyzed as a single network but this requires that teachers can nominate teachers in other schools and in these situations, the data collection process should highlight these differences. Similarly, the network model should reflect that these ties all come from a single model.

In the case of multiple types of ties or longitudinal networks, authors should explicitly state that the nodes are the same. Plotting nodes in the same orientation is highly recommended to ease confusion for the reader. At the same time, summary statistics should be computed separately because the networks are in fact separate from each other, but presented together in the Results section since the nodes are the same. For example, a single table may list each node and show the corresponding node-level statistics at each time or for each type of tie.

Finally, the type of analysis should align with the type of network data is collected. Ignoring additional dependence among networks will impact inference. For example, longitudinal models should be used for longitudinal networks, and multiplex models should be used for networks with several different types of tie.

9. Fitted Models

Depending on the type of statistical model employed, parameter estimation may need to be displayed in different ways. Regardless of how parameter estimates are summarized, appropriate intervals or error estimates should be included. For example, model-fitting software/algorithms may provide point estimates and standard errors or posterior distributions. In the latter case, posterior mean, median or modes should be reported along with some measure of variability such as posterior standard deviation or 95% credible interval.

Some models such as exponential random graph models include specific network structures in the models to account for tie dependence and parameter estimates of these structures should be included in the model. This not only reminds the reader about the specific model used, but these structures may be important for covariate interpretation.

Other models incorporate the use of latent variables as a way to account for network tie dependence. In this case, the estimates of the latent variables need not be included. For example, latent space models include estimates of latent space positions and error terms but in most substantive analyses, the positions are irrelevant to the study and used as a diagnostic tool instead. Alternatively, some stochastic blockmodels estimate cluster membership and these parameters are likely to be of substantive interest and should be included in the results.

10. Correct Model Inference and Connection Back to Theory

Many social network models appear similar to generalized linear models, which allow for similarities in parameter interpretation as well as similarities in precautions. For example, covariate parameters estimates are usually discussed as being positive or negative in nature but the ability to compare effects across models depends on the type of model. Suppose a network model estimates the effects of being in the same grade on friendship ties. A positive effect indicates that students in the same grade are more likely to have ties and a negative effect indicates the opposite. Large negative and positive effects are of course relevant and important, but the actual values of these coefficients are difficult to compare across models and close to impossible across different classes of model.

If authors do wish to discuss the values of parameter estimates, we recommend that authors examine estimates in terms of tie probabilities. For example, an effect of same grade could be 2 in one model and 3 in another model, but the overall increase in probability could be quite similar depending on other parameters in the model. This same approach applies to probit network models or other types of link functions less common in the literature. Despite these caveats, social network models lend themselves easily to substantive interpretation. Covariates can be compared in that one variable may contribute more or less to tie probabilities than others and these differences are likely important to the hypothesized theory or research question.

Furthermore, because these models are less common in the social sciences, authors may be tempted to spend additional information describing the interpretation of these models. Reviewers should be flexible to some degree to enable readership but encourage authors to maintain focus on the research.

11. Limitations of Models and Analysis

Because of the variety of analyses and models employed in social network analysis, it is extremely important that authors address the limitations of the methods used. For example, one limitation is

that social network data are often collected by survey so that networks are constructed via self-report. There may be some bias depending on the individuals surveyed or type of relationship recorded.

Other limitations include how the data were processed. It is very common for authors to dichotomize their network data, but dichotomizing data is a limitation since potentially valuable information is ignored in the analysis. Furthermore, complete case data is often analyzed and depending on the response rates, this could be a major limitation of the inference in the study. For example, if the authors only have network information from 50% of the individuals, any inference made from these individuals may not be generalizable to the population at large.

Finally, two potential limitations come from using social network models. The first is the lack of goodness of fit and model selection methods available for many other methods. Thus, some inference from these models may vary based on how the model is specified. The second idea is that homophily and influence are often confounded. Recall, influence is the idea that individuals who share ties influence each other's nodal covariates to become more similar, and homophily is the concept that individuals form ties with individuals who are similar. Most models focus on estimating effects only for homophily, but authors who are using models to estimate influence may not be able to disentangle these two phenomena.

References

- Handcock, M. S. (2003). *Assessing degeneracy in statistical models of social networks*. Working Paper 39. Seattle, WA: Center for Statistics and the Social Sciences, University of Washington. Retrieved from www.csss.washington.edu/Papers.
- Hunter, D. R., Goodreau, S. M., & Handcock, M. S. (2008). Goodness of fit for social network models. *Journal of the American Statistical Association*, 103, 248–258.
- Snijders, T. A., Steglich, C. E., Schweinberger, M., & Huisman, M. (2008). *Manual for Siena version 3.2*. Groningen, The Netherlands: ICS, Department of Sociology, University of Groningen.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.

33

Structural Equation Modeling

Ralph O. Mueller and Gregory R. Hancock

Structural equation modeling (SEM) represents a theory-driven data analytical approach for the evaluation of *a priori* specified hypotheses about causal relations among measured and/or latent variables. Such hypotheses may be expressed in a variety of forms, with the most common being *measured variable path analysis* (MVPA) models, *confirmatory factor analysis* (CFA) models, and *latent variable path analysis* (LVPA) models. For analyzing models of these as well as more complex types, SEM is not viewed as a mere statistical technique but rather as an analytical process involving model conceptualization, parameter identification and estimation, data-model fit assessment, and potential model re-specification. Ultimately, this process allows for the assessment of fit between (typically) correlational data, obtained from experimental or non-experimental research, and one or more competing causal theories specified *a priori*; most common SEM applications are *not* designed for exploratory purposes. Software packages such as AMOS, EQS, lavaan, LISREL, Mx, and Mplus are utilized to complete the computational, but not the substantive aspects of the overall SEM process. For contemporary treatments of SEM we recommend texts by Byrne (1998, 2006, 2012, 2016), Kline (2016), and Loehlin and Beaujean (2017), or, for more advanced readers, books by Bollen (1989), Hancock and Mueller (2013), Hoyle (2012), and Kaplan (2008). Specific desiderata for applied studies that utilize SEM are presented in Table 33.1 and explicated subsequently.

1. Substantive Theories and Structural Equation Models

Early in a manuscript, each model under investigation must be thoroughly justified by a synthesis of the theory thought to underlie that model. In typical SEM applications, an operationalized theory assumes the form of a *measured variable path analysis* (MVPA), *confirmatory factor analysis* (CFA), or *latent variable path analysis* (LVPA), although the analysis of more complex models (e.g., latent means, latent growth, multilevel, or mixture models) has become more popular in the applied behavioral and social science literatures. Regardless of model type, there must be strong consonance between each model and the underlying theory, as a lack thereof can undermine the modeling process. SEM's main strength lies in its ability to help evaluate *a priori* theories, not to generate them post hoc (equally, not to evaluate theories derived through prior exploration of the same data, for example with an exploratory factor analysis). Often, the articulation, justification, and testing of competing alternative models strengthens a study as it provides a more complete picture of the

Table 33.1 Desiderata for Structural Equation Modeling.

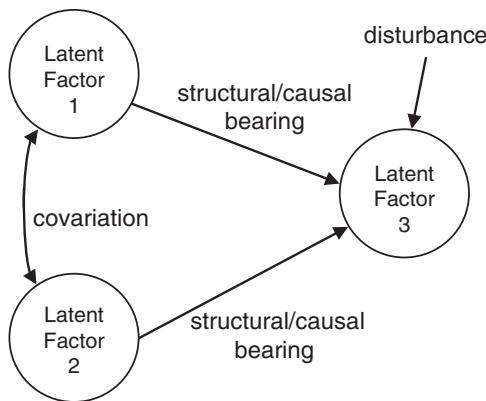
<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. Substantive theories that led to the model(s) being investigated are synthesized; a set of <i>a priori</i> specified competing models is generally preferred.	I
2. Path diagrams are presented to facilitate the understanding of the conceptual model(s) and the specification of the statistical model(s).	I
3. If applicable, latent factors are defined and their status as latent (vs. emergent) is justified.	I, M
4. Measured variables are defined and, if applicable, their appropriateness as indicator variables of associated factors is justified.	M
5. Latent factors are indicated by a sufficient number of appropriate measured variables; how the latent factors are given scale within the model(s) is addressed.	M
6. How theoretically relevant control variables are integrated into the model is explained.	M
7. Sampling method(s) and sample size(s) are explicated and justified.	M
8. The treatment of missing data and outliers is addressed.	M, R
9. The name and version of the utilized software package is reported; the parameter estimation method is justified and its underlying assumptions are addressed.	M, R
10. Problems with model convergence, offending estimates, and/or model identification are reported and discussed.	R
11. Summary statistics of measured variables are presented; information on how to gain access to the data is provided.	R
12. For models involving structural relations among latent variables, a two-phase (measurement, structural) analysis process is followed and summarized.	R
13. Recommended data-model fit indices from multiple classes are presented and evaluated using literature-based criteria.	R
14. For competing models, comparisons are made using statistical tests (for nested models) or information criteria (for non-nested models).	R
15. For any post-hoc model re-specification, theoretical and statistical justifications are provided.	R
16. Latent factor quality is addressed in terms of validity and reliability.	R
17. Standardized and unstandardized parameter estimates together with information regarding their statistical significance are provided; R^2 values for key structural outcomes are presented.	R, D
18. Appropriate language regarding model tenability and structural relations is used.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

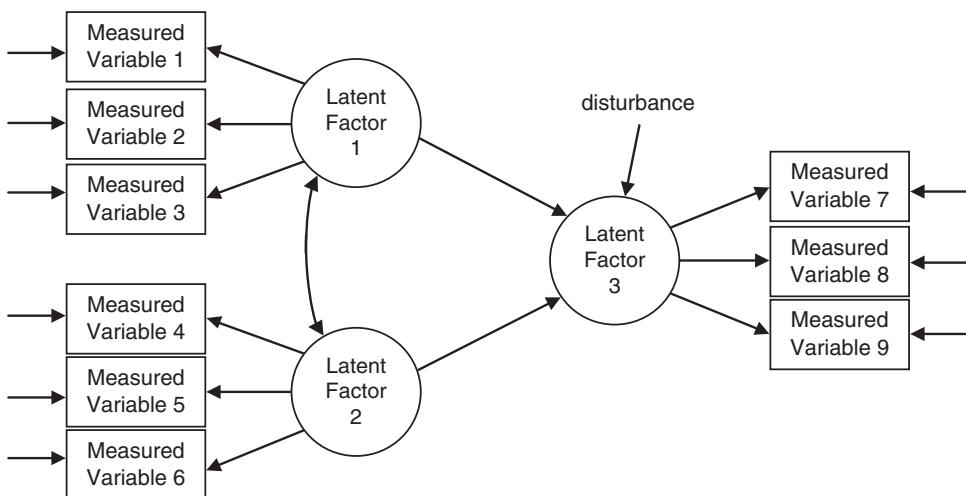
current thinking in a particular field (also see Desideratum 14). Thus, authors must convey a firm overall sense of what theoretical and/or prior empirical evidence has led to the initial conceptualization of the model(s) under study.

2. Path Diagrams

A *path diagram* is a graphical depiction of a theory relating measured and possibly latent variables and is helpful not only in representing the conceptual links among those elements but also in the specification of the statistical model. By convention, *measured variables* are represented by rectangles/squares and *unobserved factors* are expressed by ellipses/circles (and, in the case of mean structure models, occasionally a triangle is used to facilitate the modeling of means and intercept terms). Directional (one-headed) arrows point from hypothesized *causes* to *effects*, while non-directional (two-headed) arrows represent covariation between elements of a model (or a variance of an element, when the non-directional arrow returns directly to the element of origin) where the origin of that (co)variation exists outside of the model.



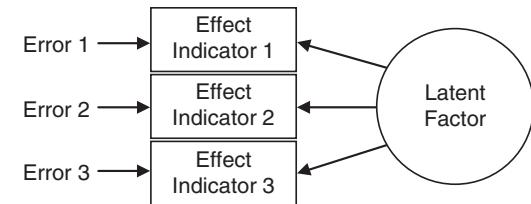
(a). Structural Model without Measurement Portion



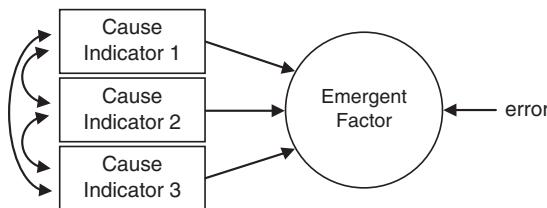
(b). Structural Model with Measurement Portion

Figure 33.1 Path Diagrams.

In investigations proposing models that involve structural relations among factors (whether those factors are *latent* or *emergent*; see Desideratum 3), a distinction should be made between the *measurement* and *structural* phases of the modeling process (see Desideratum 12). Often, the *structural portion* of such a model is the focus of the study as it represents the main theory to be tested; thus, a path diagram of the structural portion should be presented and justified early in a manuscript (see Figure 33.1a). The model may be depicted with the associated *measurement portion* that focuses on how latent variables are manifested in observable data, that is, it explicates the hypothesized relations between latent factors and their chosen measured indicators. A model with both structural and measurement portions is shown in Figure 33.1b (assuming three indicators per factor for simplicity), although details of the measurement model might be better presented and explained in the Methods section of a manuscript where specifics about the instrumentation are typically discussed. In sum, authors are encouraged to utilize appropriately detailed path diagrams to complement and illustrate their written explanations of the theory being tested.



(a). Latent Factor with Effect Indicators



(b). Emergent Factor with Cause Indicators

Figure 33.2 Latent vs. Emergent Factors.

3. Latent vs. Emergent Factors

SEM, in its common forms of confirmatory factor analysis (CFA) and latent variable path analysis (LVPA), addresses theory-based relations among latent variables (factors). Here, the term *latent* adds meaning to its popular definition of being unobserved or unobservable: it connotes a factor hypothesized to have a causal bearing on one or more measured indicator variables. This causal relation implies that the latent factor explains variance in its measured indicator variables and induces covariance among them. That is, individuals vary on the measured indicators in part because they vary on the underlying factor; their indicator scores covary because they have a common latent cause. SEM analyses involving factors typically assume them to be latent, and researchers must explain and justify that the factors are indeed causal agents of their observable *effect* indicators (see Figure 33.2a).

On the other hand, theory and/or the nature of the available data occasionally dictate that measured variables serve as *cause* indicators of constructs therefore described as *emergent* (rather than effect indicators of constructs considered latent). Variation in such measured cause indicators (which themselves might or might not covary) is now hypothesized to cause and partially explain variation in the emergent factor (see Figure 33.2b). For example, the construct *socioeconomic status* (SES)—measured by indicators such as *parental income*, *education*, and *occupational prestige*—should be modeled as emergent, not latent: It seems to make little theoretical sense, for instance, to propose that SES is the cause of parents' education level. More sensible would be an emergent system, whereby changes in one's SES are caused (at least in part) by changes in parental income, education, and occupational prestige. The analysis of models involving emergent constructs is certainly possible but generally more challenging (for example, parameter estimation difficulties can occur) than models involving latent variables only. It is incumbent upon the researcher to justify *each* factor's status as latent (or emergent) rather than simply presume a latent status for all factors. The latter could lead to a misspecified measurement model and, in turn, to incorrect inferences regarding the relations between the construct in question and other variables (latent or measured) within the model. For a more detailed discussion on latent vs. emergent variable systems, consult Kline (2013).

4. Measured Variables

All measured variables in a model should be defined clearly, or if they are previously established measures, an accessible reference to their description should be provided. Measured variables may be *exogenous* (independent; having no causal inputs) or *endogenous* (dependent; having one or more causal inputs) within a model. Each may serve as (a) an effect indicator of a latent factor (where it is theoretically clear that the factor has a causal bearing on the measured variable), (b) more rarely, a cause indicator of an emergent factor (where it is theoretically clear that the variable has a causal bearing on the latent factor; see Desideratum 3); or (c) a stand-alone variable, not serving as an indicator of any factor in the model (e.g., sex of respondent).

Measured variables can be individual items from larger scales or scores obtained by aggregating across several response items. Exogenous variables may be dichotomous (two categories) or continuous without requiring any special model estimation procedures. The analysis of ordinal variables with fewer than five scale points (exogenous or endogenous, including effect indicators of latent factors) typically requires special estimation procedures and/or input data consisting of alternate measures of association (e.g., polychoric correlations; see Chapter 5, this volume, and Finney & DiStefano, 2013). Other measured variable types that might warrant accommodation, or a defense as to why no such accommodation was made, include count variables and *censored* variables (e.g., an income variable with a top category of \$50,000 and up).

5. Indicator Variables of Latent Factors

Without context, latent factors may have any number of measured variable indicators, but in order to determine if an adequate number was utilized in a particular application, three key issues should be considered. First, overall model identification (the ability to estimate all model parameters) can be helped or hindered by the chosen number of indicators per latent factor. With three or four indicators, the factor does not require the estimation of more parameters (e.g., loadings, error variances) than the data supplied from its measured indicators. Having fewer than three indicators is certainly possible and occasionally even necessary due to limited availability. Using a single indicator variable of a factor may be accomplished by setting the error variance of the indicator variable to zero (effectively equating the variable with the factor) or to some portion of the measured indicator's variance, depending upon known reliability of the indicator variable (e.g., a reliability coefficient of .80 for a measured variable implies that 20% of its variance is residual, unexplained by its latent factor). Using two indicator variables for a latent factor typically requires no fixing of error variances to known constants as long as the factor relates to one or more other factors in the model. This case does run the risk, however, of creating an *empirical under-identification* problem should the relation of the factor to others within the model be estimated as zero or near zero.

Second, having three or more indicators tends to enhance the quality of the construct (its theoretical breadth as well as its ability to replicate across samples, that is, construct validity and reliability). Contrary to early methodological literature, having additional indicators does not appear to burden estimation even for relatively modest sample sizes. At some point, however, diminishing returns are expected. Our experience is that having four to six indicators of reasonable quality (e.g., standardized loadings exceeding .6 or .7 in absolute value) is practically ideal, although more indicators can be feasible as well.

Third and finally, as factors are latent, they have no inherent *metric*. Hence, they must be assigned units within the model. For an exogenous factor, this can be accomplished by either fixing its variance (typically to 1, giving the factor standardized units) or by keeping the variance freely estimated and instead fixing a path to one of its indicator variables (typically to 1, giving the factor the metric of the indicator variable). For an endogenous factor, assigning units is typically accomplished by the latter method, although some software programs allow the specification of the former as well. Researchers should be explicit in how latent factors were scaled and if their software package handled the assignment automatically.

6. Control Variables

Researchers might examine parameter estimates in a model of theoretical interest after controlling for the effects of specific variables, thus ensuring that the estimates are above and beyond the linear effects of external elements (including background variables capturing individual characteristics, such as IQ, and/or environmental conditions, such as family income or parental education). Although some applied researchers have attempted such control by *a priori* partialing the variables from the data, the practice of analyzing such *residualized* data with SEM is questionable. Instead, incorporating the control variables directly into the model is the preferred strategy. Typically all control variables (measured and/or latent) are allowed to covary as exogenous predictors within the model. Each has a structural path to all exogenous and endogenous factors (and stand-alone variables) within the structural, but not the measurement, portion of the model. Interpretation of the paths within the structural portion proceeds normally, now acknowledging that they have been purged of the linear effects of the control variables.

Two additional points of elaboration are worthwhile. First, control variables should not generally be modeled as indicators (cause or effect indicators) of a single factor (e.g., a demographic factor) unless indeed the factor makes sense as a continuum in its own right. Second, the type of control described here is *linear*, not nonlinear or multiplicative. Thus, for example, if a researcher believes the structural paths of interest actually differ as a function of the control variables (i.e., that the control variables *moderate* the structural relations), then a model with multiplicative predictor terms or, in some cases, a multisample analysis (see Chapter 34, this volume) should be considered. Either way, the researcher should clarify the expected nature of the control variables' relations with the structural elements of the theoretical portion of the model, and how the modeling strategy employed exacts the proper type of control.

7. Sampling Method and Sampling Size

Sampling methods (e.g., random, stratified, cluster) must be made explicit in the text. Stratified sampling techniques typically yield sampling weights (see Chapter 35, this volume) whose purpose, when applied to the sample data, is to weigh individual cases more or less in order to maximize the sample's representativeness of the target population. In such cases researchers, should delineate how these weights were incorporated into their modeling. Multistage (cluster) sampling approaches, where groups of individuals are sampled at a time, will yield a *hierarchical* (nested) structure to the data (e.g., students within classrooms within schools). This introduces some dependence among observations in the sample, thereby violating the assumption of independence of observations. In this case researchers should utilize a multilevel SEM approach or present evidence of sufficiently small effect of the dependence among clusters of cases (e.g., a small design effect) to justify not pursuing the more complex multilevel approach (e.g., Stapleton, 2013).

In order to determine the sample size required for a SEM analysis, researchers should consider both adequacy for correct parameter estimation and for desired level of statistical power. A common guideline for obtaining trustworthy *maximum likelihood* (ML) estimates is to have at least five cases per model parameter (not per variable); when employing other estimation methods that require less stringent distributional assumptions and/or that are tailored specifically for ordinal data (e.g., *asymptotically distribution free*; *arbitrary generalized least squares*; *weighted least squares*; *Satorra-Bentler rescaling corrections*), more cases are generally needed. If the available sample size is substantially smaller, researchers must provide justification from the SEM methodological literature (e.g., sample size necessary for sound parameter estimation can be reduced by having latent factors of high quality; see Desideratum 16). Regarding power, sample size should be justified on two fronts. First, authors should explain how the available sample size provides adequate power for relevant tests of the data-model fit as a whole (e.g., using the confidence interval for the *root mean square error of approximation*,

RMSEA; see Desideratum 13). Second, the sample size should provide sufficient power to detect individual model relations of key theoretical interest (e.g., structural connections among latent factors). For a detailed discussion of statistical power in SEM, consult Hancock and French (2013).

8. Missing Data and Outliers

Within the context of SEM, classic missing data techniques (e.g., listwise or pairwise deletion, mean substitution) are generally considered inadequate unless the amount of missing data is so small as to be trivial. Methods currently considered acceptable include *full-information maximum likelihood* (FIML) estimation and *multiple imputation* (MI), but their availability varies across SEM software packages (for details, see Enders, 2013). The former implicitly allows all subgroups of individuals, defined by differing patterns of available data, to contribute to those parameters' estimation, which their data are able to inform. The latter, on the other hand, imputes values for those missing but does so multiple times to determine average parameter estimates from across several possible sets of partially imputed data. Both methods assume that data are *missing at random*, that is, that the missing data mechanism for each variable is independent of that variable (e.g., people failing to provide income information has nothing to do with their income). Authors must identify and justify which missing data algorithm was utilized (FIML or MI) and report the proportions of cases missing for each applicable variable.

With regard to cases that were not missing but were otherwise aberrant, criteria for the detection and possible removal of such outliers from further analysis must be addressed. To start, if descriptive background data led the researcher to conclude that specific cases were inappropriate for generalizing to the population of interest (e.g., some data came from foreign students in otherwise English-speaking classrooms), such preliminary (but post hoc) exclusion criteria should be made explicit. From a univariate statistical perspective, outliers can be determined by comparing cases' standard (z) scores against some reasonable threshold, such as ± 3 ; cases exceeding that threshold could be removed. Multivariate outliers (where a case's scores on individual variables may not be unusual but its combination of scores sets it apart) may be diagnosed using Mahalanobis distance (D) or squared distance (D^2). Because D^2 values follow a χ^2 distribution in large samples, each case has an associated p -value; cases with extreme values, say $p < .001$, could be removed as outliers. Finally, cases may be evaluated in terms of their influence on the overall multivariate kurtosis (to which normal-theory estimation methods such as maximum likelihood are particularly sensitive); criteria for cases' removal are relative to the metric of kurtosis employed by the software. Thus, with regard to outliers, authors are responsible for detailing the criteria used for outlier removal or for addressing other methods automatically employed by software packages (such as down-weighting extreme cases).

9. Software and Estimation Method

Typical structural equation models may be estimated using a wide variety of software packages, including, but not limited to, AMOS, EQS, lavaan, LISREL, Mplus, Mx, and SAS PROC CALIS. These packages tend to produce very similar if not identical results for most common applications. As modeling needs become more demanding, however, either in terms of addressing assumption violations (e.g., non-normality; categorical data) or model complexity (e.g., multilevel models), packages differ in their capabilities and/or the manner in which they meet such needs. Authors should report which software package was used and the specific version of that software, as these packages are constantly evolving. If a package was chosen to meet specific modeling needs, an explanation thereof is warranted.

Regarding specific estimation methods, although maximum likelihood (ML) is a default in virtually all packages, it should not be chosen without an understanding of its limitations (e.g., inflated parameter z -values and model χ^2 values under non-normality), an explicit rationale for its selection, and data-based justification thereof. Summary statistics detailing a lack of skewness (e.g., <2) and

kurtosis (e.g., <7) for measured variables can help assuage concerns regarding non-normality at a univariate level. Addressing more directly the multivariate assumption underlying normal-theory methods (e.g., ML; *Generalized Least Squares*), Mardia's *normalized coefficient of multivariate kurtosis* should also be reported. Although no universal guideline exists, values around 3 or less can be reassuring. By extension, authors utilizing an alternate estimation procedure (e.g., asymptotically distribution free; weighted least squares) or rescaling procedure (e.g., Satorra–Bentler corrections) should fully justify the selection, including a discussion of its assumptions (or specific lack thereof) and its applicability to the data and model(s) at hand.

10. Problems with Convergence, Estimates, and Identification

In practice, SEM analyses seldom proceed smoothly. Whether by programming error or uncooperative data, problems involving convergence, estimation, and/or identification inevitably occur. The estimation process may fail to converge within a default number of iterations; linear dependencies may occur that prevent some parameters' identification and hence estimation; offending estimates such as negative error variances (Heywood cases) may arise; matrices may be reported as non-positive-definite; and so forth. As SEM is a process, documenting the steps followed is necessary just as in other scientific endeavors. To facilitate model convergence, perhaps it was necessary to specify start values other than the default (typically derived from *two-stage least squares*), or some variables needed to be rescaled to make their variances less extreme relative to others. Perhaps an error variance needed to be constrained (e.g., to zero or slightly above) to eliminate a Heywood case, or a ridge regression correction was necessary to address a non-positive-definite covariance matrix. Regardless of the issues faced, whether those described above or others, authors are responsible (not unlike a bench scientist keeping a lab notebook) for detailing challenges that arose and the corrective actions taken, or for stating explicitly that no such problems were encountered.

11. Data Display and Accessibility

In order to facilitate verification of study results, and allow for exploration of competing explanations for those results, the American Psychological Association (APA) requires that in their manuscripts, authors present "informationally adequate statistics." Although, clearly, not all journals containing SEM analyses are governed by APA, the rationale for this requirement is sound and is herein endorsed. Authors should provide adequate summary information for readers to be able to verify the presented results (e.g., sample size, covariance matrices or correlation matrices with standard deviations, and possibly other relevant statistics such as means and reliabilities). In instances where the number of variables modeled prohibits tabling summary information economically, information as to how to acquire such information should be provided (e.g., from a website). Where advanced estimation methods are used that draw directly from raw data and cannot be accomplished using summary statistics, authors should provide information as to how readers may gain access to the data.

12. Two-Phase Modeling Approach

For the analysis of latent variable path models, a two-phase modeling process is generally recommended to facilitate the diagnosis and potential remediation of data-model misfit. In the first or *measurement* phase, the model is temporarily re-specified such that all latent variables are allowed to freely covary (along with stand-alone and control variables). If the data satisfactorily fit this measurement model (see Desideratum 13), the second phase can commence; if not, re-specification of the measurement model may be entertained. Such modifications are typically informed by theoretical

considerations, *Lagrange multiplier tests (modification indices)*, and/or relatively large residuals, and most often take the form of either error covariances or cross-loadings (paths from factors to secondary measured indicators) not already in the model; see Desideratum 15. If satisfactory data-model fit was not achieved even after such modification(s), further modeling should be terminated. If, however, reasonable fit is achieved, the second phase of modeling is entered.

In the second or *structural* phase, the originally hypothesized structural relations are re-inserted among the factors (and stand-alone and control variables) while preserving whatever measurement model modifications might have been made during the first phase. Poor data-model fit for this initial structural model can no longer be due to the measurement portion of the model but to the structural portion. At this point, both statistical and practical evaluation of the overall model should occur. Because this initial structural model is nested within the final measurement model, a χ^2 difference test should be conducted to assess the statistical difference between the two (see Desideratum 14). Ideally, fit would not degrade statistically significantly, however a fit difference in the final measurement model and initial structural model is typically expected since structural perfection is unlikely. If the data do not fit the initial structural model to a satisfactory degree, it should be rejected and authors should explicitly acknowledge so. If, on the other hand, authors re-specified the conceptual structure, strong statistical and theoretical justifications must be provided and the now exploratory nature of the analysis acknowledged (see Desideratum 15).

13. Data-Model Fit

A central issue addressed by any SEM analysis is the assessment of the fit between observed data and the hypothesized model (for a now classic overview, see Bollen & Long, 1993). While a χ^2 test is commonly reported for this purpose, it is viewed by most as overly strict given its power to detect even trivial deviations of the data from the proposed model. Researchers should therefore report multiple fit indices, typically drawing from three broad classes (while many indices are available, only commonly recommended ones are listed together with literature-based target values to retain a model):

- *Absolute indices* evaluate the overall discrepancy between observed and implied covariance matrices (and possibly means); fit improves as more parameters are added to the model: the *standardized root mean square residual* (SRMR) should typically fall below .08.
- *Parsimonious indices* evaluate the overall discrepancy between observed and implied covariance matrices (and possibly means) while taking into account a model's complexity; fit improves as more parameters are added to the model, as long as those parameters are making a useful contribution: the *root mean square error of approximation* (RMSEA) and its associated 90% confidence interval should fall below .08, and preferably below .05.
- *Incremental indices* assess absolute or parsimonious fit relative to a baseline model, usually the null/independence model (which specifies no relations among observed variables): the *Normed Fit Index* (NFI), *Non-normed Fit Index* (NNFI; also referred to as *Tucker-Lewis Index*, TLI), and/or *Comparative Fit Index* (CFI) have .95 as a desired minimum value.

Because absolute, parsimonious, and incremental data-model fit indices can lead to inconsistent conclusions, reviewers should insist on fit results from different classes. If, after considering several indices, data-model fit is deemed acceptable (and judged best compared to competing models, if applicable), the model is retained as tenable and individual parameter estimates may be interpreted. If, however, evidence suggests unacceptable data-model fit, it *might* be appropriate to re-specify the model (see Desideratum 15) to improve fit.

14. Model Comparisons

In as much as authors follow the recommendation to propose and test competing theories in their investigations (see Desideratum 1), they are then obligated to offer comparative judgments regarding the relative tenability of these alternative explanations of what gave rise to the observed data. These judgments can be based on statistical and/or descriptive evidence, depending upon the *nested* (hierarchical) nature of the structural equation models. When two models are nested (such as, but not limited to, when the estimated parameters in the former are a proper subset of those associated with the latter), fit comparisons can be accomplished with a formal χ^2 difference test (also referred to as a *likelihood ratio test*) or rescaled χ^2 difference test if using rescaled/robust statistics. For example, an orthogonal CFA model (factors are specified to be independent) is nested within an oblique one (factors are allowed to covary), and data will fit the former less well than the latter as evidenced by a larger model χ^2 . However, if the χ^2 difference test (or rescaled difference test) indicates no statistically significant difference in fit, a researcher focusing on model parsimony might prefer an orthogonal over an oblique explanation as no statistical distinction has been established.

When two models do not have a nested relation, researchers must compare the models descriptively, often using information criteria such as the *Akaike information criterion* (AIC) or *Bayesian Information Criterion* (BIC), which express the expected ability of models to cross-validate: Models associated with smaller information criterion values are preferred over models with larger information criterion values. Descriptive comparisons of other data-model fit indices are also possible to discern practical differences in fit, but irrespective of comparison approach, authors must be clear that these results are relative, not absolute: while data might be judged to fit one model better than another, both models might exhibit unsatisfactory data-model fit as gauged by indices designed to evaluate each model individually (see Desideratum 13).

15. Model Re-specification

In a strict sense, *any* hypothesized model is, at best, only an approximation to reality; the remaining question is one of degree of that misspecification. With regard to *external* specification errors—when irrelevant variables were included in the model or substantively important ones were left out—remediation can only occur by re-specifying the model based on additional relevant theory. On the other hand, *internal* specification errors—when unimportant paths among variables were included or when important paths were omitted—can potentially be diagnosed and remedied using *Wald statistics* (predicted increase in χ^2 if a previously estimated parameter were fixed to some known value, e.g., zero) and *Lagrange multiplier statistics* (also referred to as *modification indices*; estimated decrease in χ^2 if a previously fixed parameter were to be estimated). As these tests' recommendations are directly motivated by the data and not by theoretical considerations, any resulting re-specifications must be acknowledged as data-driven and exploratory in nature and might not lead to a model that resembles reality any more closely than the one(s) initially conceptualized.

For example, a statistically nonsignificant path could merely be the result of insufficient statistical power, and its removal from a model could become theoretically misleading. We believe authors should leave nonsignificant paths in the model, as it preserves the originally theorized model while still communicating that, within the model's context, a hypothesized relation did not establish itself beyond chance. On the other hand, implementing suggestions from modification indices to add paths to the measurement or structural portion of a model could in fact merely be a capitalization on random covariation in the current sample and not be indicative of true population relations. In the case of measurement model modifications (e.g., cross-loadings, error covariances), statistical support (e.g., change in χ^2 values) should be supplemented by a theoretical rationale for each parameter addition. For structural model modifications, such as adding paths among factors not originally hypothesized, statistical and theoretical support should be strengthened by an explicit

discussion addressing the tension between the original hypotheses and the exploratory and apparently contrary re-specifications.

16. Validity and Reliability of Latent Factors

Each latent factor should be evaluated in terms of its *validity*, that is, its ability to represent the construct it is hypothesized to represent, as well as its *reliability*, that is, the ability to replicate across new sets of data. With regard to the former, researchers should address the extent to which the factor relates to elements it should, and does not relate to elements it should not. Those elements first and foremost are the factor's own effect indicators, for which standardized and/or unstandardized loadings and significance test results should be reported. Researchers should observe patterns of loadings that are relatively high for measured variables expected to reflect the factor, and relatively low (ideally zero) for variables intended to reflect other factors. One validity index commonly recommended is the *variance extracted* (the average squared standardized loading for a factor's indicators) with target values around .50 and above. Another gauge of a factor's validity worth addressing is the degree to which its relation to other factors matches theoretical expectations.

With regard to reliability, while in theory a factor is perfectly reliable as it is an error-free entity, in practice a factor would not be expected to replicate perfectly should the same individuals provide new scores on the factor's indicators (assuming no recollection of the prior measurements). Thus, reliability reflects the extent to which a factor is expected to replicate (i.e., correlate with itself), given that it is indicated by data containing error. Although some researchers report Cronbach's α based on the sum of a factor's indicators (often after standardization), this index is inappropriate as it reflects reliability of a composite rather than the reliability of the factor as reflected in its measured indicators. Thus, we recommend instead that authors report maximal reliability (*Coefficient H*; e.g., Hancock & Mueller, 2001) for each factor, as it is an estimate of the correlation that a factor is expected to have with itself over repeated administrations. Values above .70, and preferably higher, would be considered desirable in order to establish the replicability of each hypothesized factor.

17. Results for Specific Structural Equations

For either measured or latent variable path analysis models, once the overall data-model fit has been assessed and deemed satisfactory (see Desideratum 13) and, if applicable, evidence of the quality of latent variables has been presented (e.g., construct validity and reliability; see Desideratum 16), more detailed results regarding the structural relations of interest should be offered. Individual parameter estimates (i.e., the *direct* structural effects from one variable to another) should be listed in standardized and unstandardized form to facilitate comparisons to results obtained in subsequent studies (though technically, one set of estimates might be acceptable, given that it is derivable from the other if sufficient information is provided). Statistical significance information for key parameters is essential, in the form of symbolic designations (e.g., asterisks) or actual test statistics (e.g., *z*-values), and authors should present coefficients of determination (R^2 values) for each measured or latent outcome of theoretical interest. Depending on the study's purposes, a presentation of the *indirect* and *total* structural effects, along with their statistical significance information (preferably obtained through bootstrapping), might also prove useful in understanding and interpreting associations among structurally related measured and/or latent variables. Note, however, that SEM software packages do not typically indicate the statistical significance of individual indirect effects but only of the compound indirect effect via all potentially intervening mediators, and of the overall total structural effect of one variable on another.

18. Interpretative Language

Readers should be wary of authors' claims that acceptable data-model fit implies a model was "confirmed" or that a particular theory was proven to be "true," especially after post-hoc re-specifications (see Desideratum 15). Such statements are grossly misleading given the typically nonexperimental nature of the data's origins, and given that alternative, structurally different yet mathematically equivalent, models always exist that would produce identical fit results and thus would explain the data equally well. At most, a model with acceptable fit may be interpreted as *one* tenable explanation for the associations observed in the data.

Following satisfactory data-model fit, the interpretation of individual parameter estimates is permitted to involve explicit causal language, *as long as this is done from within the context of the particular causal theory proposed* and the possibility/probability of alternative explanations is raised unequivocally. Though some might disagree, we think that explicit causal statements are more honest than implicit ones and are more useful in articulating the study's practical implications within the guiding theoretical framework (for more on the role of causality in SEM and other statistical analyses, consult Pearl, 2009). In the end, SEM is a powerful tool at the researcher's disposal for testing and interpreting theoretically derived causal hypotheses from within an a priori specified causal system of measured and/or latent variables. However, we urge reviewers to continually remind authors to resist the apparently still popular belief that the main goal of SEM is to achieve satisfactory data-model fit results; rather, it is just to get one step closer to the truth. If it is true that a proposed model does not reasonably approximate reality, then reaching a conclusion of *misfit* between data and model should be a desirable goal, not one to be avoided by careless re-specifications until satisfactory levels of fit are achieved.

References

- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bollen, K. A., & Long, J. S. (1993). *Testing structural equation models*. Newbury Park, CA: Sage.
- Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS*. Mahwah, NJ: Lawrence Erlbaum.
- Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Byrne, B. M. (2012). *Structural equation modeling with Mplus*. New York: Taylor & Francis.
- Byrne, B. M. (2016). *Structural equation modeling with AMOS* (3rd ed.). New York: Taylor & Francis.
- Enders, C. K. (2013). Analyzing structural models with missing data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) (pp. 493–519). Charlotte, NC: Information Age Publishing.
- Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) (pp. 439–492). Charlotte, NC: Information Age Publishing.
- Hancock, G. R., & French, B. (2013). Power analysis in covariance structure modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) (pp. 117–159). Charlotte, NC: Information Age Publishing.
- Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future—A Festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International.
- Hancock, G. R., & Mueller, R. O. (Eds.) (2013). *Structural equation modeling: A second course* (2nd ed.). Charlotte, NC: Information Age Publishing.
- Hoyle, R. H. (Ed.) (2012). *Handbook of structural equation modeling*. New York: Guilford Press.
- Kaplan, D. (2008). *Structural equation modeling: Foundations and extensions* (2nd ed.). Thousand Oaks, CA: Sage.
- Kline, R. B. (2013). Reverse arrow dynamics: Feedback loops and formative measurement. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) (pp. 41–79). Charlotte, NC: Information Age Publishing.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford Press.
- Loehlin, J. C., & Beaujean, A. A. (2017). *Latent variable models* (5th ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge: Cambridge University Press.
- Stapleton, L. M. (2013). Multilevel structural equation modeling with complex sample data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) (pp. 521–562). Charlotte, NC: Information Age Publishing.

34

Structural Equation Modeling *Multisample Covariance and Mean Structures*

Richard G. Lomax

This chapter represents an extension to Chapter 33 in this volume on basic structural equation modeling (SEM). Next to consider is the development and testing of theoretical models in two more advanced contexts. First, multiple sample SEM (MS-SEM) considers the invariance (i.e., equality) of parameters across populations (i.e., equivalence of covariance structures). For example, Jöreskog and Sörbom (1993, example 10) tested whether the same factor structure of the SAT verbal and math sections is present in two samples of students (e.g., by examining invariance of factor loadings, the factor correlation, and the measurement error variances). Second, structured means SEM (SM-SEM) additionally assesses mean differences between populations (i.e., intercept parameters). For example, Jöreskog and Sörbom (1993, example 13) examined mean differences between academic and non-academic boys on the latent variable of verbal ability at grades 5 and 7.

The hypothesized models can be tested utilizing SEM software such as AMOS (Arbuckle, 2013), EQS (Bentler, 2014), LISREL (Jöreskog & Sörbom, 2015), or Mplus (Muthén & Muthén, 2015), and with data from experimental, quasi-experimental, cross-sectional, or longitudinal studies. For full-length descriptions of basic SEM analyses, including some discussion of MS-SEM and SM-SEM, we recommend the following textbooks: Blunch (2013), Byrne (2006, 2009, 2012, 2014), Kline (2016), Schumacker and Lomax (2016), and Wang and Wang (2012). Examples of multiple sample and structured means models are described in the SEM software manuals, in various chapters of Hancock and Mueller (2013) and Hoyle (2012), in Lomax (2013), as well as in articles such as Aiken, Stein, and Bentler (1994), LARRC Consortium (2015), Lomax (1983, 1985), Morris et al. (2012), and Shumow and Lomax (2002).

Multiple sample structural equation modeling (MS-SEM) is appropriate for testing theoretical model(s) across various subsamples (e.g., by age, ethnicity, SES, gender, grade). Here the researcher wants to know whether a particular model holds for each of these subsamples, or whether there are some differences among the subsamples for various parameters (e.g., factor loadings, structure coefficients). This indicates whether a particular theoretical model applies rather broadly across contexts or not.

Structured means structural equation modeling (SM-SEM) is utilized to examine mean difference (or intercept) parameters. Such questions may seek to determine if there is a mean difference in an SES latent variable for public versus private school populations (Lomax, 1985), or whether

Table 34.1 Desiderata for Structural Equation Modeling: Multisample Covariance and Mean Structures.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. Justification for theoretical model(s) to be tested, including hypothesized differences across samples, is made based on the available theory and research in the substantive area.	I
2. Path diagram(s) are presented to display the theoretical model(s) to be tested, including noting sample differences in figures.	I
3. Populations and samples are each described, including sampling method(s) and sample sizes.	M
4. The measured variables that are used to indicate the latent factors are fully described (e.g., scales, as well as psychometric properties for each sample).	M
5. Name and version of the software utilized is reported. In addition, the method of parameter estimation is discussed and justified based on the scales and distributions of the observed variables, and assumptions are assessed.	M, R
6. Methods of treating missing data and outliers are described.	M, R
7. Table(s) of correlations, means, and standard deviations for each sample are presented, and access to raw data is facilitated, if applicable.	M, R
8. Problems with identification, convergence, non-positive definite matrices, and inadmissible solutions are reported and resolved.	R
9. Various global goodness-of-fit indices are reported for each model (and for each sample as appropriate), and chi-square difference and other relevant tests are reported to compare nested models.	R
10. Parameter estimates and statistical significance are reported for each sample in tables, path diagrams, and/or text.	R
11. Model modifications (if any) are reported, including theoretical and statistical justifications.	R
12. Results are discussed in the context of assessing the invariance of the theoretical model(s) tested, as well as limitations noted from prior issues.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

there is a mean difference in a student achievement latent variable for these same two populations, controlling for SES (Lomax, 1985).

Table 34.1 indicates the desiderata for MS-SEM and SM-SEM around which the remainder of this chapter is organized.

1. Justification for Theoretical Model

In reading an SEM application, it is often difficult to determine exactly where the theoretical model(s) tested come from. Here the concern is mostly with the structural model, that is, the model where the relations among the latent variables are hypothesized. Reviewers often come to the Results section of an SEM study and suddenly theoretical model(s) appear, seemingly out of thin air. Where exactly do these theoretical models come from? Often, the reviewer has no knowledge of the basis of the models, just that they are presented as a path diagram in one or more figures of the Results section.

The theoretical model(s) should be developed from the available theory and research in the substantive area being studied. This is the major purpose of the literature review in research utilizing SEM. More specifically, the literature review should compile the research on each of the specified relations among the latent variables (i.e., the structure coefficients). Each structure coefficient is

essentially a mini-hypothesis of the relation between two latent variables. Evidence could be drawn from whatever types of research are available (e.g., correlational, experimental, non-experimental, qualitative), as well as theoretical work. After reading the literature review, the reviewer should be presented with the theoretical model(s) to be tested and knowledge about where the model(s) came from. In short, the literature review should guide the reviewer from the available literature to the theoretical models to be tested.

For MS-SEM and SM-SEM, the literature review should also lead up to whether the researcher expects to find group differences in the theoretical model(s). For example, the research might suggest that male and female adolescents develop differently for specified constructs. This would then lead to hypotheses that certain parameters in the SEM model(s) will behave differently by gender. In MS-SEM we can test whether specific parameters appear to be invariant (i.e., the same) across populations or noninvariant (i.e., different) across populations. In SM-SEM we can also test whether there appear to be latent mean differences across the populations. As an example, in the Lomax (1985) model of schooling, a series of MS-SEM models determined that the factor loadings and structure coefficients appeared invariant for the public and private school populations. In subsequent analyses, an SM-SEM model inferred that there were several latent mean differences associated with school type, such as for constructs relating to home background, academic orientation, and extra-curricular activity.

2. Path Diagrams

A path diagram should be included for each theoretical model tested. Such path diagrams can be presented in at least two different forms depending on their intended purpose. One purpose of a path diagram is to present the theoretical model as part of the literature review. This conceptual diagram will not include any results or estimates but merely displays the latent variables and how they are theorized to relate to one another. In other words, this type of path diagram describes the theoretical structural model only. The conceptual path diagram can either be depicted at the beginning of the literature review for framing purposes, or at the end of the literature review for summative purposes. For multisample SEM, hypothesized differences can also be depicted in the path diagram (e.g., F > M to hypothesize a path for females having a stronger relation than for males; or one can use different line types, such solid or dotted, to denote paths expected to be different for each sample). The measurement model is implied by the description and/or list of observed variables and which latent variables they assess (see Desideratum 4). Details of the observed variables are contained in the Methods section.

A second purpose of a path diagram is to display the results of the analysis. In other words, key parameter estimates (perhaps even including standard errors, or z values, or statistical significance, in parentheses) are shown in the diagram. Thus, in a single diagram the most important results of a particular model can be shown. Unless the model is quite small and simple in nature, only the structural model is usually depicted in this type of path diagram. In other words, the measurement model is not included unless there is sufficient space in the figure. In multiple sample SEM, the estimates for each sample should be shown, either in tables (columns for each sample) or in figures (path diagrams displaying the structure coefficients for each sample, such as .52/.48 for males and females, respectively).

3. Sampling Issues

In most SEM studies, the only information about sampling provided is sample size, which is not adequate to assess the quality of the work. There is additional information about three issues that the reviewer should be made aware of in terms of the sampling frame of the study. This information

allows for judging the sampling adequacy of the sample as well as the populations the results can be generalized to.

First, information about the definition of the populations must be included (e.g., males and females, aged 13–17 in the US, who are attending public schools and receiving regular education). Second, the type of sampling procedure utilized for each population needs to be described (e.g., random, cluster, stratified, convenience). That is, authors should describe how these particular sample observations were selected from their respective populations (e.g., a strategy of stratified random sampling by ethnicity was followed to ensure adequate representation of different ethnic groups, or if a particular under-represented group was oversampled to increase representation). And finally, precisely how the data were collected from these samples needs some attention (e.g., assessments were individually administered to subjects over two days by a trained specialist); that is, details need to be presented on how the data were actually collected from each of the samples. Information about these sampling issues will also be useful when discussing missing data (see Desideratum 6) and deciding upon an appropriate method of estimation (Desideratum 5).

4. Latent Variables and Observed Variables

A key portion of the Methods section for any SEM study is the selection and preparation of the observed variables (for examples, see LARRC Consortium, 2015; Morris et al., 2012; Shumow & Lomax, 2002). There are five measurement issues that need to be considered. First, for each latent variable or factor, one or more observed variables are utilized. One method for dealing with this is in the form of a table listing the latent variables and each of their associated observed measures. Alternatively, the measurement of the latent variables through the observed variables can be described in the text (e.g., a section for each latent variable, with a subsection for each observed variable). This information helps to define the measurement model in terms of latent variables and observed variables. Second, the types of measures actually being used should be described. Are the observed variables individual items, composites of items (i.e., unweighted or weighted sums of individual items in a subscale, such as item parcels), or factor scores? In other words, the type of measures being utilized should be discussed.

Third, what are the psychometric properties of the observed variables? Except for the most commonly used and well-known measures, information should be given on the reliability and validity of each observed variable. This information needs to include some basic evidence through reliability and/or validity coefficients, which often can be documented in a sentence or two for each observed variable. In multisample SEM, psychometric evidence should also be presented for each population being assessed (as some measures could behave differently in different contexts).

Fourth, it would be useful for the reviewer to have some information about the possible scale values (e.g., minimum/maximum or range), type of measurement scale (i.e., nominal, ordinal, interval, ratio), the number of categories for ordinal measures, and whether any rescaling or recoding of the variable values has been undertaken. Oftentimes the scale of an original variable has been altered to fit the needs of the individual study or analysis (e.g., to rescale an observed variable when there is a negatively worded item; or to generate a more normally distributed observed variable through the use of a statistical transformation, see Desideratum 5). Consideration of multiple samples should also be made.

Finally, how are the observed variables distributed for each sample? In other words, are the measures reasonably normally distributed, what do the distributions look like in terms of skewness and kurtosis (univariate and multivariate), and how is any non-normality taken into account (e.g., statistical transformations; or the use of an estimation procedure more robust to non-normality)? This information is also particularly relevant for the selection of an appropriate method of estimation, as discussed next (see Desideratum 5).

5. Software and Method of Estimation

There are a number of software packages available for conducting multisample SEM analyses (e.g., AMOS, EQS, LISREL, Mplus), as well as R-based software (e.g., lavaan, OpenMx). Thus, the name and version of the software utilized should be reported. All of the packages are somewhat comparable, and thus no overarching recommendation can be made. However, each program does have certain features that distinguish it from the others. As well, because the field of SEM has changed so rapidly over the past two decades, utilizing a recent version of the software is necessary.

As in any quantitative research study involving the use of inferential statistics, applicable statistical assumptions must be evaluated and potential violations dealt with in an appropriate manner. In other words, statistical assumptions form part of the foundation for any inferential statistic. Without reasonably meeting those assumptions we cannot count on results having much value.

In multisample SEM studies, most methods of estimation make some assumption about the distribution of the observed variables. Maximum likelihood estimation is the most commonly used method of estimation in SEM and assumes that the variables are multivariate normal. Thus, it is crucial to determine whether the data meet this distributional assumption for each sample. If evidence (e.g., skewness, kurtosis, statistical test of normality) suggests that the data are not reasonably multivariate normally distributed, then maximum likelihood may not be appropriate (note that the evidence may not be consistent across samples). Here the offending observed variables could be (a) transformed into a new variable that is somewhat better behaved, (b) deleted if there are a sufficient number of observed variables to assess each latent variable, or (c) a method of estimation more robust to non-normality be utilized instead of maximum likelihood (e.g., generalized least squares, weighted least squares, diagonally weighted least squares). It should be noted that this only addresses univariate distributional issues; two or more variables would have to be considered simultaneously to address multivariate distributional issues. Any such decisions need to be applied across all of the samples.

Maximum likelihood is typically recommended unless the data deviate substantially from multivariate normality and/or include categorical variables (e.g., Lei & Lomax, 2005). Thus, the choice of a method of estimation needs to be tied to the distributional properties of the data, as well as to the measurement scales of the variables, all of which need to be reported. In other words, an SEM study needs to describe the distribution of the data, the method of estimation, and the measurement scales of the observed variables. In the Shumow and Lomax (2001) model of school safety, the observed variables were fairly-well normally distributed in a univariate sense (e.g., in terms of skewness and kurtosis) in each of the ethnic group samples, thus maximum likelihood was selected as the method of estimation.

6. Treatment of Missing Data and Outliers

In multisample SEM studies, little information is typically provided about the presence and treatment of missing data for each sample. This is curious because rarely does a social or behavioral scientist have a dataset with absolutely no missing data, and having multiple samples only increases the likelihood of missing data. The inclusion of sampling framework information can easily lead into a discussion of how missing data were treated. Were only complete cases used (i.e., listwise deletion), were complete cases included for each pair of observed variables (i.e., pairwise deletion, where different pairs of variables could have different sample sizes), was a missing data replacement method utilized (e.g., mean imputation, similar response pattern imputation, multiple imputation such as expectation-maximization or Markov chain Monte Carlo), or was a method used that works around the missing data (i.e., does not impute the missing data, but works with the available data, such as full-information maximum likelihood)? Excellent references on missing data include Enders (2010) and McKnight, McKnight, Sidani and Figueiredo (2007).

A description of how the missing data were treated and a justification for that method should be included, in addition to the initial and final sample sizes. For example, Arbuckle and Wothke (1999, example 17) considered a dataset with a sample of 73 girls on six psychological tests. Approximately 27% of the data were missing, complete data were only available for seven cases, and thus full-information maximum likelihood was selected as the method for dealing with the missing data. In a multisample SEM study, there would be cause for concern if the percentage of missing data differed widely across samples. This would suggest that the data were perhaps missing systematically rather than randomly (e.g., due to an intervention or some other effect). It would also suggest less confidence in results for samples with larger percentages of missing data and perhaps even compromise sample invariance or sample mean difference tests.

Another data treatment issue deals with outliers. An outlier is an observation that is quite different from the rest (e.g., often defined in a univariate sense as more than 2 or 3 standard deviations beyond the mean). Outliers can cause any of the following issues: normality problems; improper solutions (see Desideratum 8); correlations being reduced in strength or magnitude (i.e., closer to zero); as well as other serious issues. Thus, it is always important to report the detection of outliers for each sample.

Outliers can be a function of any of the following situations: (a) a malfunctioning instrument (e.g., a computerized testing or tape recorder problem); (b) a data recording error (e.g., recorded a 60 instead of a 6 when the data were gathered); (c) a data entry error (e.g., typed a 60 into the data file instead of a 6 when the data were entered); (d) an error in observation (e.g., observed and coded an incorrect behavior); (e) an inappropriate use of administration instructions (e.g., gave subjects more time than the directions called for); or (f) an accurate observation (i.e., a true outlier). Obviously errors should be corrected whenever possible. Otherwise, a rationale should be made for the treatment of outliers, and the number of outliers deleted should be reported for each sample.

7. Data Table for Samples

Once the data have been fully prepared, then the analysis can proceed accordingly. An important initial piece of information for the reviewer is a summary table of correlations, means (for SM-SEM), and standard deviations (or variances) for the set of observed variables. This is typically presented through a table in matrix form. In the case of multiple samples, either (a) separate tables can be shown for each sample, or (b) pairs of samples can be given in one table with one sample being shown above the main diagonal of the matrix and a second sample below the main diagonal (e.g., see tables 2 and 3 in Shumow & Lomax, 2002).

The summary table allows the reviewer and others to examine the data from a descriptive perspective and even to conduct their own analyses (i.e., either to verify the results or to test additional models). The complete matrix should be given in the data table for each sample, unless there are too many observed variables to include for practical purposes. In this case the table should be made available by some other means (e.g., website, e-mail). In summary, there should be sufficient information in any SEM application to allow the reviewer (and reader) to be able to replicate the results (e.g., theoretical model, sampling information, measures, software and method of estimation, data tables, and results).

8. Problems in Obtaining a Proper Final Solution

Reviewers need to be aware that a proper or final solution does not result 100% of the time, even with the specification of a theoretically justifiable model. For example, a negative variance estimate (i.e., a Heywood case), or correlation estimates that exceed 1.0, are not proper solutions. In addition,

most methods of estimation utilized in SEM are iterative, meaning that numerous iterations occur prior to reaching a final solution. When a final solution is not generated, absurd estimates (e.g., a standardized factor loading of 100) can occur and should not be taken seriously. These issues are more likely to occur with multisample SEM. Thus reviewers need to (a) determine if any of these issues have been reported and dealt with, and (b) question any such results, as they cannot be trusted.

9. Global Goodness-of-Fit Indices and Model Comparisons

An essential part of the SEM results that must be reported are the global goodness-of-fit indices. That is, a well-fitting model is necessary prior to the reporting of parameter estimates for that final model. Simulation studies of fit indices have been conducted yielding the following two recommendations for reviewers: (a) the chi-square measure can be greatly influenced by several characteristics, including sample size, model complexity, and non-normality, and thus should never be used as the sole measure of model fit; and (b) no other single fit measure has been shown to be “best” in all contexts. In short, it is always recommended that multiple fit indices be reported in SEM studies, typically three to five measures. For single sample SEM applications, considerable attention is devoted to fit indices in Chapter 33 of this volume (Desideratum 13).

For multisample SEM applications, the chi-square difference test is one fit index that is always recommended in order to examine a series of nested models (i.e., one model is nested within another model when the variables are the same, but one or more parameters are different). For example, to compare an initial model with no parameters invariant across populations to a second model with only factor loadings invariant across populations, a difference in chi-square statistic can show the extent to which the fit of the second model has deteriorated. If the chi-square value does not decrease by a statistically significant amount, then the factor loadings have been shown to be statistically invariant across the samples tested; as such we would retain a hypothesis of factor loading invariance across populations. If the chi-square value does decrease by a statistically significant amount, then the factor loadings have been shown to be statistically different across the samples tested; we would infer some degree of factor loading non-invariance across populations. Partial invariance of a subset of factor loadings can also be evaluated in a similar fashion. Thus, the chi-square difference test can help assess whether the factor loadings are the same or different across populations. The invariance of other types of parameters (e.g., structure coefficients) can also be evaluated by comparing two nested models.

Reviewers should note that other fit indices can also be utilized in this context. For example, Cheung and Rensvold (2002) suggested that changes in fit indices such as CFI or GFI are useful for assessing measurement invariance. Such discussions typically are only relevant for continuous variables (although Bovaird & Koziol, 2012, considered ordered-categorical indicators). Meredith (1993), among others, discussed different models for measurement invariance (e.g., strict, strong, weak), although this is beyond the scope of this chapter.

In addition, to determine the strength of any effect in SM-SEM, an effect size measure can also be computed. For example, a standardized effect size measure can be determined where the intercept of an equation for the second group is divided by the square root of the disturbance or error variance of that equation (e.g., Hancock, 2001). These effect sizes can then be judged in accordance with guidelines such as those provided by Cohen (1988).

To be more specific about the conduct of MS-SEM, a prescribed series of nested models is typically evaluated. Model 1 is one in which none of the parameters are invariant across populations. As an example using gender, this would mean that none of the parameters in the female population model are constrained to be equal to their respective parameters in the male population model.

Model 2 constrains the factor loadings to be the same across the populations, but not other model parameters (i.e., parameters other than the factor loadings are free to vary across populations). If the fit according to the chi-square difference test substantially deteriorates from Model 1 to Model 2, then the factor loadings are statistically different for the samples and we infer some degree of non-invariance across populations. If the fit does not deteriorate substantially, then the factor loadings are not statistically different for the samples, and there is support for population invariance. When factor loadings are invariant, then each latent variable represents the same entity across populations, and subsequent tests of other parameters make sense.

Model 3 considers whether the measurement error variances are invariant (which occurs only on rare occasions and is not expected or necessary). Model 4 assesses whether the latent independent variable variances and covariances are invariant. Model 5 is used to determine whether the structure coefficients are invariant. Finally, Model 6 examines whether the structure or prediction error variances are invariant (also rare). It should be noted that the sequence will depend on (a) the specific parameters in the model, (b) the specific hypotheses to be tested, and (c) the results of the previously tested models.

The idea is to test a series of nested models to determine which sets of parameters are invariant and which are not. Thus, MS-SEM is a method for determining whether the same covariance structure is operating across the populations. In addition, tests can be done at the individual parameter level (i.e., partial invariance). For example, one can assess whether *one* particular factor loading is invariant across populations as opposed to *all* factor loadings being invariant. A table of results should be provided that lists each model tested, including the chi-square value (with degrees of freedom and *p* values reported), chi-square difference tests, and some indication of the statistical significance for each chi-square difference test (e.g., see table 6 of Lomax, 1985).

10. Parameter Estimates and Statistical Significance

Another important part of the Results section is a complete reporting of the estimates, standard errors, and statistical significance for every individual free parameter (e.g., through a *z* value, and/or some sort of notation regarding statistical significance). With MS-SEM, this means that all of the results need to be reported for each sample. An exception would be where particular parameters are found to be invariant across samples; here each sample would have the same results for those parameters. With multiple samples, the rows of the table can be the individual parameter estimates and the columns can be the different samples (e.g., tables 4 and 5 in Shumow & Lomax, 2002). For SM-SEM, the estimates of the mean difference parameters will also need to be reported (e.g., either in the bottom portion of the results table, such as table 7 of Lomax, 1985, or alternatively in a separate table just of structured means results, such as table 2 of Lomax, 1983).

For ease of reviewing these tables, it is best that the results be presented using observed and latent variable names, rather than Greek symbols, and that the different types of parameters be clearly labeled (e.g., factor loadings, measurement error variances, structure coefficients). It is also recommended that the unstandardized results be reported because it is those estimates that are being tested for invariance in MS-SEM and SM-SEM. As previously described under Desideratum 2, results can also be reported in the form of a path diagram.

11. Model Modification

Model modification is one of the most poorly reported aspects in SEM studies. When the initial model does not have adequate model fit, which is typical, then modifications to that model are considered. Reviewers need to assess three questions in the area of model modification. First, were relevant parameters statistically different from zero (typically evaluated through *z* values), at some

nominal alpha level, and in the expected direction (see also Desideratum 9)? For example, the fit of the theoretical model is described as acceptable, but no information is provided about the significance of the estimates. As well, model fit may be quite strong, but the estimates do not make sense, or are significant in the wrong direction, or are not even significant. Thus, an adequate model fit means little if the parameter estimates do not support the hypothesized model.

Second, what information was utilized in considering model modification (e.g., substantively from theory and prior research; or statistically from residuals, modification indices, expected parameter change statistics, Lagrange multiplier statistics)? As it is rare that is such information provided, even a brief description is useful. Third, how was the initial theoretical model modified? For example, were new relations added or non-significant relations trimmed from the original model? In most applications only the results of the final model are presented. But reviewers should know that testing a single model is rarely the case. In the interest of journal space, while only the final model results should be fully reported, a brief overview of the models tested and the modifications made should be included.

In multisample applications, model modification becomes a bit more complex. The initial step is to arrive at one, final model that fits across all samples when tested separately, prior to any multiple sample analysis. In other words, any issues have already been dealt with (those that relate model fit, estimation and modification, as previously described) so that this is the model to go forward with. Next a baseline multisample model is tested where there are no constraints across samples. From there additional constraints are made for the invariance of particular parameters, as previously described. For example, in looking at invariance of factor loadings, meaning the latent variables across samples constitute the same entity, invariance could range from no invariance (none of the factor loadings are invariant), to partial invariance (some of the factor loadings are invariant), to full invariance (all of the factor loadings are invariant). Thus the model modification process for multisample situations involves an examination of the amount of invariance for relevant types of parameters (e.g., factor loadings, structure coefficients).

12. Discussion of Results in Context of Theoretical Model(s)

One of the main purposes of the Discussion section of an SEM study is to relate the results of the theoretical model(s) tested to the literature previously reviewed. The following questions could be considered in the Discussion. What evidence is there that a particular model has been supported for one or more populations? Are there some portions of the model that fit the data rather well and others that are not? Should additional latent variables be included and/or other latent variables eliminated based on the analysis? Do we need to refine our list of observed variables in terms of seeking out higher quality measures? Do some of the paths need to be eliminated? What differences were detected across samples? These are just some of the issues that a Discussion section can address.

In addition,, the study of multiple samples allows a form of model validation as to whether a particular model fits well for different samples (e.g., for males and females), or in different contexts (e.g., for different countries or for different cohorts). The Discussion section could also describe future methods for validating a model or models in other samples or contexts. For example, does the same model of schooling apply in the United States, Estonia, and Japan?

References

- Aiken, L. S., Stein, J. A., & Bentler, P. M. (1994). Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. *Journal of Consulting and Clinical Psychology*, 62, 488–499.
- Arbuckle, J. L. (2013). *IBM SPSS Amos 22 user's guide*. Chicago, IL: SPSS.

- Arbuckle, J. L., & Wothke, W. (1999). *Amos 4.0 user's guide*. Chicago, IL: SPSS.
- Bentler, P. M. (2014). *EQS (Version 6.2)*. Encino, CA: Multivariate Software.
- Blunch, J. J. (2013). *Introduction to structural equation modeling using IBM SPSS Statistics and AMOS*. Thousand Oaks, CA: Sage.
- Bovaird, J. A., & Koziol, N. A. (2012). Measurement models for ordered-categorical indicators. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 495–511). New York: Guilford.
- Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming*. New York: Routledge.
- Byrne, B. M. (2009). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. New York: Routledge.
- Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge.
- Byrne, B. M. (2014). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. New York: Routledge.
- Cheung, R. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indices for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9, 233–255.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.
- Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, 66, 373–388.
- Hancock, G. R., & Mueller, R. O. (2013). *Structural equation modeling: A second course* (2nd ed.). Charlotte, NC: Information Age.
- Hoyle, R. H. (2012). *Handbook of structural equation modeling*. New York: Guilford.
- Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago, IL: Scientific Software International.
- Jöreskog, K. G., & Sörbom, D. (2015). *LISREL (Version 9.2)*. Skokie, IL: Scientific Software International.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford.
- LARRC Consortium (2015). Learning to read: Should we keep things simple? *Reading Research Quarterly*, 50, 151–169.
- Lei, M., & Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Quarterly*, 12, 1–27.
- Lomax, R. G. (1983). A guide to multiple sample structural equation modeling. *Behavior Research Methods and Instrumentation*, 15, 580–584.
- Lomax, R. G. (1985). A structural model of public and private schools. *Journal of Experimental Education*, 53, 216–226.
- Lomax, R. G. (2013). Structural equation modeling. In Y. Petscher & C. Schatschneider (Eds.), *Applied quantitative analysis in the social sciences* (pp. 245–264). New York: Routledge.
- McKnight, P. E., McKnight, K. M., Sidani, S., & Figueiredo, A. J. (2007). *Missing data: A gentle introduction*. New York: Guilford.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525–543.
- Morris, D., Trathen, W., Lomax, R. G., Perney, J., Kucan, L., Frye, E. M., Bloodgood, J. W., Ward, D., & Schlagal, R. (2012). Modeling aspects of print-processing skill: Implications for reading assessment. *Reading and Writing: An Interdisciplinary Journal*, 25, 189–215.
- Muthén, B. O., & Muthén, L. K. (2015). *Mplus (Version 7)*. Los Angeles, CA: Muthén & Muthén.
- Schumacker, R. E., & Lomax, R. G. (2016). *A beginner's guide to structural equation modeling* (4th ed.). New York: Routledge.
- Shumow, L., & Lomax, R. G. (2001). Predicting perceptions of school safety. *The School Community Journal*, 11, 93–112.
- Shumow, L., & Lomax, R. G. (2002). Parental efficacy: Predictor of parenting behavior and adolescent outcomes. *Parenting: Science and Practice*, 2, 127–150.
- Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. Chichester: Wiley.

35

Survey Sampling, Administration, and Analysis

Laura M. Stapleton

The use of surveys in social science research is abundant and may appear straightforward, but can involve a complex set of procedures. *Total survey error* refers to all of the errors that could occur in researchers' attempts to gain valid information about a population with the use of surveys. These errors have been placed into five categories: coverage error, nonresponse error, editing and processing errors, measurement error, and sampling errors (Groves, 1989). *Coverage error* refers to the failure to provide some members of the population the chance to be selected into the sample, *nonresponse error* refers to the failure to obtain responses from all members of the selected sample, *editing and process errors* refer to the failure to capture data accurately from the respondents, *measurement error* represents the failure of the observed response to reflect the true opinion of the sample member, and *sampling error* refers to the fact that sample statistics are not expected to exactly reflect population parameters.

Methods to acknowledge and address each of these types of errors are covered in this chapter and should be, to the extent possible, addressed in research manuscripts. Specifically, this chapter addresses three main areas of the survey process: appropriate methods of sampling from a specified population, developing survey items and administering the survey, and undertaking appropriate analyses based on the sampling design. The sampling design defines how one obtains a sample intended to be representative of the population to which researchers would like their results to generalize. Sampling is discussed in Lohr (1999, 2008), and for the more advanced reader, Kalton (1983) and Kish (1965). Development of survey items, and the appropriate methods of administering the survey in order to obtain high response rates and quality responses, is multifaceted and development strategies depend on the topic to be measured and the target population of the survey. An excellent, comprehensive resource with practical guidelines is provided by Dillman, Smyth, and Christian (2014) and Forsyth and Lessler (1991). More theoretical treatment regarding the cognitive response process is provided by Tourangeau, Rips, and Rasinski (2000). Finally, analysis of data is straightforward if a simple random sample has been taken (analysis covered in most behavioral and social science textbooks assume such a sampling strategy). However, many surveys are not conducted with a simple random sampling technique and thus special analytic procedures must be undertaken and these procedures might include sampling weights, and strata and/or cluster indicators. Analyses undertaken using such sampling design elements are detailed at a basic

Table 35.1 Desiderata for Survey Sampling, Administration, and Analysis.

<i>Desideratum</i>	<i>Manuscript Section(s)*</i>
1. The population for generalization is described and justified.	I, D
2. The survey or questionnaire development process is described and includes a discussion of the evaluation of item validity.	M
3. The sampling frame is defined and the procedures for obtaining it are described.	M
4. The type of sampling process utilized is explained (such as multistage, random, systematic) and defended as to why it is appropriate for use with the target population.	M
5. The survey administration method is outlined including the mode (such as face-to-face, phone interview, web survey, or mailed survey), use of incentives, and number of contacts. The administration process is defended as to why it is appropriate for use with the particular sample and questionnaire content.	M
6. The response rate is provided and discussed.	R
7. An analysis of possible sources of non-response is conducted and the possibility of the creation of post-stratification weights is addressed.	M, R
8. The data analysis includes components to address any disproportionate selection probabilities (or non-response adjustments).	M, R
9. If probability sampling other than simple random sampling is used, the estimated design effects of the means of key variables are provided.	R
10. The analysis includes an appropriate method to adjust for any dependencies resulting from multistage sampling or efficiencies gained from stratified sampling.	M, R
11. A discussion of the limitations with regard to questionnaire item validity, sampling strategy, response rate and analysis decisions is provided.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

level in Lohr (1999) and Stapleton (2008, 2013), and at a more advanced level in texts by Heeringa, West, and Berglund (2010), and Kish (1965). A simple list of best practices for the entire survey research process is provided by the American Association for Public Opinion Research (2008) and a volume edited by de Leeuw, Hox, and Dillman (2008) contains accessible chapters that relate to every step of the survey process. Specific desiderata for studies that incorporate the use of surveys for collection of behavioral and social science data are presented in Table 35.1 and are described in detail in the following sections.

1. Population for Generalization

Although the term *survey* has been used to connote different things across people and disciplines, at its heart, the term should imply that a questionnaire has been administered to a sample of members from a given population to which researchers would like to generalize; a questionnaire that is administered to all members of a given population is considered to be a census and not a survey. Therefore, it is imperative that in an applied study involving the administration of a survey, authors specify the population to which results are intended to generalize. It is from this target population that the sampling frame is developed (see Desideratum 3). There are times when researchers would like to generalize to a large population, such as all US adults aged 18 to 65, but the sampling procedures do not allow such a broad generalization; for example, if a telephone survey is administered based on random digit dialing, the actual population from which the sample is drawn only includes adults with a cellular or land-line telephone. This narrowed population, then, is referred to as the

sampling frame population. If there is a discrepancy between the target population and sampling frame, authors should carefully describe the members of the target population who will be excluded by sampling from only the narrower sampling frame, in terms of both numbers and characteristics. To the extent that the members of the target population and sampling frame differ in their behaviors and attitudes measured on the survey, this discrepancy is termed coverage error.

2. Questionnaire Development Process

The questionnaire development process is one that is often poorly defined in research manuscripts, if not totally ignored. It can be argued, however, that it is one of the most important aspects of survey research; if validity of the measures obtained from the survey cannot be assured then any subsequent data analysis and interpretation of parameter estimates is questionable. Given research findings with regard to the effect of item wording on response quality, it is essential that authors describe the process undertaken to develop and evaluate the questionnaire. If using a previously developed questionnaire, authors should cite a reference to the development and validation information for that questionnaire. If using a newly developed questionnaire, authors should provide a description of how the items were generated and whether items were reviewed by outside parties. Additionally, a discussion should be provided of the pilot testing process and the results shared from reliability assessments (both internal consistency for scales and test-retest reliability of specific items and scale scores, as appropriate) and validity assessments (such as cognitive interviewing results or statistical estimates of relations with external criteria).

Specific issues in item development that have been found to affect response quality and thus induce measurement error and that should be considered in questionnaire development include, for rating scale measures, the *valence* of the item stem (whether the stem is phrased negatively or positively), the number of response options, and whether anchors are provided for those options. Although surveys often include the “strongly agree” to “strongly disagree” Likert format, this type of measurement has been found to be problematic for two reasons: without extreme statements there usually is limited variability in the obtained responses, and the responses are difficult to interpret across individuals due to the vague quantification of response options.

Not only the wording and structure of the actual items are of concern but the format of the questionnaire has been found to affect response rate and response quality. Dillman and colleagues (2014) provided excellent suggestions for the actual page layout and construction. Authors should provide a copy of the questionnaire used or provide an internet link or citation to allow readers to evaluate the survey for possible order or questionnaire construction effects, as well as item wording effects.

3. Sampling Frame

To obtain a representative sample of a given population of interest, a sampling frame needs to be determined. As specified in Desideratum 1, given a target population of interest, the researcher must then attempt to identify a list of all members of that population from which to sample. If a list is indeed available (e.g., all students enrolled at a particular university of interest) then the concept of a sampling frame is fairly simple (although the definition of “enrolled at the university” on a specific date would need to be determined and reported). Often, however, no list of members of the target population exists; in this case a sampling frame might have to be pieced together from separate sources (and duplicate and missing units identified and addressed), or the sampling frame might have to be built as part of the sampling process. For example, if researchers were interested in a target population of all undergraduate students at 4-year US colleges and universities, they

would be quick to note that no such list currently exists. However, each of the 2,000+ colleges and universities in the population of interest does have a list of its own students. Therefore, the sampling frame might exist in theory and can only be practically accessed once specific higher level units (the institutions in this example) are contacted. Whether the list already exists or must be constructed as sampling takes place, authors should delineate the process of building the sampling frame from which the final sample was determined.

4. Sampling Design

There are several types of sampling designs that are appropriate for generalization to a target population. A few sampling strategies exist that might be less appropriate and are discussed briefly at the end of this section. For generalization, especially with correlational research designs, it is best to have what is referred to as a *probability sample* where each member of the sampling frame has a known probability of being selected into the sample.

One probability sampling scheme is *simple random sampling* (SRS) in which each unit in the sampling frame has an equal chance of selection. SRS can only be achieved if the researcher has a list of the entire sampling frame (either physically or theoretically as in random digit dialing whereby a computer can generate all possible combinations of numbers). Another SRS-equivalent sampling strategy is referred to as *systematic sampling*. In this process, a sampling interval is determined to arrive at a final sample number. For example, if there are 5,000 students at a university and a sample of 200 students is desired, then a list of the students sorted in random order can be created and every $5,000/200 = 25$ th student on this list is selected for the sample. Note that each population member's probability of selection (π) is the inverse of this interval (e.g., $\pi = 1/25 = .04$). Systematic sampling can also be used when a list is not available, by selecting, for example, every 10th shopper to enter a store or every 20th voter to arrive at a polling place. In order to use this systematic sampling strategy, an estimated size of the population coming to the location would be needed, along with a desired sample size; these two values would provide the appropriate sampling interval.

Another probability sampling design is termed *stratified sampling*. In this sampling method, members of the sampling frame are split into mutually exclusive categories (strata) and then elements are sampled from each category to ensure specific representation of members of each stratum; the selection rate might or might not be proportional to the population size within each stratum depending on the researchers' needs. The resulting sample will be more representative based on the strata as compared to the use of SRS and systematic sampling, but this result has implications for assumptions underlying sampling error estimates as will be discussed in Desideratum 10.

Multistage sampling designs refer to the selection of *primary sampling units* (PSUs) as a first stage of sampling (such as the random selection of US colleges and universities) and then sampling of one or more lower level units within each of the selected PSUs (such as faculty or students within those selected universities). Multistage sampling is usually undertaken when a list of units in the population is not available or it is more efficient to collect data within clusters. For example, it would be easier to collect opinions of 10 students at each of 20 college and universities than to go to possibly 200 different campuses to reach 200 randomly selected students.

As part of multistage sampling, a technique called *probability proportionate to size* sampling can be used where the probability of selection of the PSU depends on the size of the PSU. If an equal number of final elements is intended to be drawn from each PSU (such as 10 students at each campus), then if PSUs are drawn with equal probability, students at small institutions will have a higher probability of being in the sample and the final sample will over-represent students at small campuses. Therefore, a sampling design to result in approximately equal probabilities of selection of lowest level units involves selecting PSUs directly proportionate to the size of the PSU, such that

small PSUs have smaller chances to be in the sample and vice versa. This technique is further discussed in Desideratum 8.

Sampling designs can incorporate any one or all of the previously mentioned strategies. For example, a stratified, multistage systematic sample might be used by researchers wanting to survey freshmen at colleges and universities in the US. First, a list of 4-year institutions would be obtained and each might be identified by sector (public/private). Researchers might specifically select institutions using SRS from each sector to guarantee that the sample would include a specified number or percentage of the sample within each sector. Once institutions are selected, and assuming the registrars were not permitted to provide lists of all enrolled students, the sample of students could be selected using stratified systematic sampling by day and time. On specific selected days and times researchers would be on campus and approach every *i*th student who passes a certain location or locations.

Because the statistical analysis approach taken depends on the sampling strategy used with probability samples, authors should report the specific sampling strategy taken and at each level if a multistage sample is drawn. Authors also should report the selection rate either overall, with SRS, or at each level of the selection process for more complex designs. If a stratified sampling design was used, the strata should be defined.

Two additional sampling strategies, both non-probability designs, are prevalent in current literature. Convenience sampling and chain-referral sampling are both seen, with the latter being a more defensible strategy under some circumstances. *Convenience sampling* is a process by which respondents are identified in a manner that is convenient to the researcher: undergraduate students at a university, patients in a given clinic, people on the street. With this sampling approach, the sample is not drawn from the full population of interest to the researcher and therefore the generalizability of any of the researcher's findings is questionable. If researchers indicate that they want to generalize findings to all US university students and then report on a sample of psychology students who participate in required subject pool research at University X, then there is a severe disconnect between the target population and the sample; coverage errors would be extremely likely. Authors who report studies based on convenience samples should defend why they believe that the findings can be generalized to the population of interest and also provide information regarding the similarity of the sample participants with target population characteristics. Even if such a similarity can be demonstrated (such as with commonly available data such as age and race/ethnicity), there are other ways that the convenience sample might not be similar to the target population. For example, subjects might differ on motivational or attitudinal characteristics not often measured. In general, convenience sampling is more acceptable with experimental designs, given that the desire usually is not to report on the finite population relations among measures but rather to examine the effect of a manipulated variable under the control of researchers.

The second non-probability sampling strategy, referred to as *chain-referrals* (including snowball sampling), has recently gained in use and opinion. With this sampling strategy, a representative sample of some given population is first found, and those individuals with the characteristic of interest are asked to refer the researcher to similar individuals. Such a strategy can be used to determine population estimates of "hidden" populations such as drug users or the homeless (Frank & Snijders, 1994). If chain-referrals are used, authors should describe the method used to obtain the initial sample and demonstrate that it obtained a fairly representative initial group and provide information about the number of waves or links in the chain referral collection.

5. Survey Administration

Because the method of survey administration has been found to influence both the survey response rate (and thus non-response error) and the responses themselves (and thus measurement error),

authors should provide information about the administration process, including the mode of administration, the number and type of contacts, and whether anonymity or confidentiality was ensured.

The most typical survey modes include paper and pencil, face-to-face interview, phone interview, and web-based administration. There is no preferred survey administration mode, as its success depends on such considerations as the content area of the survey, the target population for which the survey is intended to be used, the anonymity, the length, and the time and labor resources available to administer the survey. Researchers should take care to determine whether the chosen mode might pose a problem in terms of response rate or response bias. For example, for sensitive topics, a face-to-face interview, in which an interviewer might gain rapport with the respondent, can yield more valid data, as can an anonymous web survey. A web survey also can be efficient and inexpensive, but is subject to a specific non-response bias for those lacking access and/or familiarity with the internet. A paper and pencil mail survey tends to yield poorer response rates but is a fairly inexpensive mode and can accommodate longer surveys. Phone surveys provide quick turnaround but can be limited to shorter surveys and require a sizable labor force to administer.

In addition to the mode of survey used, authors should provide information about the survey administration process. Dillman et al. (2014) expressed many practical suggestions for obtaining high response rates and valid information from each item response. Particularly, a multi-contact system is viewed as necessary, which includes possible pre-notification that a survey is coming, the sending of the survey, and multiple non-response contacts with the final contact being of a different mode than the initial contact (for example, by phone if the initial survey was attempted to be administered by mail). An additional consideration in the survey administration process is the anonymity of the survey. If anonymous, responses have been found to have less bias under some contexts, however researchers lose the ability to track respondents and thus typically must send non-response contacts to all members of the sample or use a method whereby respondents send back a postcard indicating that they have responded. If survey response is desired to be tracked, researchers must put an identifying number or code on each survey; note that a name is not necessary.

For surveys that involve interviewer interaction with the respondent (such as with a face-to-face or telephone interview), authors should indicate the level of training that interviewers received or experience that they have and whether interviewer effects might have yielded bias in measurement (for example, having male interviewers ask female adolescents about depressive symptoms).

Survey administration also includes the process of transcribing data into a data set and this data processing step can involve error. In the case of web surveys and computer-assisted telephone interviews, the data are already entered, although authors should verify that testing of the system has occurred prior to use. For other types of surveys, the data entry process should be outlined, including methods of quality control such as double entry and random quality checking.

6. Response Rate

Response rates inform researchers about the possibility for nonresponse bias as a potential component in survey error. Although the calculation of response rates may appear simple on the surface, it is actually fairly complex. In the late 1990s a group of survey research organizations agreed upon a set of standards in calculating survey outcome rates including contact rates, refusal rates, and eligibility rates. These definitional standards were published by the American Association for Public Opinion Research (2006). To illustrate the definitional complexities in reporting response rates, consider the following example. With a telephone survey, calls are made but when there is no answer, a message on an answer machine or voicemail is *not* left. Should that household or person be considered a non-respondent? Did he or she have the opportunity to respond to the survey?

The standards are particularly helpful in determining how to report rates under complex sampling schemes. Researchers should use these reporting standards to inform their readers of the amount of non-response and thus the possibility of non-response error.

Although a common concern in survey administration is the level of response rate that is needed for appropriate inference to the population, such standards do not exist. In the late 1970s, the US Office of Management and Budget (OMB) indicated that data collection needs a minimum response rate of 75% and proposed federal projects with anticipated response rates less than 50% should not be approved. The current OMB guidelines, however, have been altered to indicate that surveys need to yield “reliable results” and indicate that a high response rate is one source of reliability (Smith, 2002).

7. Sources of Non-Response

Non-response can be at an item or a unit (person) level. Issues influencing possible non-response should be delineated. Groves and Couper (1998) grouped the influences on non-response into four categories, two of which are under the control of the researcher and two that are not. The first two include survey protocols and interviewer training as addressed in Desiderata 4 and 5. The second two include the social climate for surveys in general (sometimes referred to as “survey fatigue”) and the personal characteristics of the respondent.

As an example of the latter influence, suppose that a researcher wanted to report the average alcohol consumption of university students and finds that disproportionately more women than men responded to a survey. Such disproportionate non-response rates would likely result in non-response bias in parameter estimates. Assuming men consume more alcohol than women, the average in the obtained sample likely would be lower than the average in the population because women are over-represented in the sample due to their higher response rate.

Non-response does not necessarily indicate that estimates from the sample will be biased or inaccurately reflect population parameters. A challenge to the researcher, then, is to determine whether non-response bias exists. Several approaches can be taken to evaluate possible non-response bias, including examining disproportional response rates using sampling frame information, analyzing external data to evaluate whether the sample is distributed similarly to the population, dissecting information from interviews about possible reasons why item and/or unit non-response exists, and examining whether the number of contacts needed to convert a sample element to a responder is related to sample characteristics. Additionally, one may conduct a subsequent survey of non-responders; although difficult to implement, it can provide valuable information about why some members of the selected sample did not participate in the survey. Authors should share results from these non-response analyses with the reader.

If a possible non-response bias exists, post-stratification weighting adjustment can be used to compensate for the disproportionate non-response rate. For the alcohol consumption survey example, sampling weights for responses from men would be adjusted higher to account for their lower response rate (see Desideratum 8 for a discussion of the creation and use of weights in an analysis).

8. Sampling Weights

The sampling weight is the inverse of the selection probability, $1/\pi_i$, where π_i is the selection probability for the i th sample member, as introduced in Desideratum 4. The sampling weight can be thought of as the number of people in the population that this specific sample member is representing. If a SRS has been used in selecting the sample, then all sampling weights for observations will be equal and can be ignored in the analysis. However, if the sampling design includes stratification, multiple

stages, disproportionate selection probabilities across strata, or adjustments for non-response, then inclusion of sampling weights in the analysis typically will be necessary to obtain unbiased estimates of population parameters. Care should be taken in understanding how the selected software utilizes the sampling weights. Some software functions (such as “weight cases” in SPSS) uses weights as frequency weights and assumes that the sample size is equivalent to the sum of the weights and therefore the raw sampling weight will need to be scaled.

If not properly accounted for in the analysis, disproportionate selection of elements into the sample can adversely affect the resulting population estimates from an analysis. There are many reasons for the use of disproportionate sampling rates, and three situations that result in sampling weights that differ across sample members are discussed here: multistage sampling, over-sampling by stratum, and post-stratification adjustments. First, consider the example of a simple two-stage sampling design. Suppose we want to obtain a sample of 5,000 freshmen from 4-year colleges and universities in the US, and that there are 2,000,000 freshmen in about 2,000 institutions in the population. One way to obtain the desired sample would be to randomly sample 5,000 students out of the total pool of 2,000,000 students. To do this, we would use a selection probability of $\pi = \frac{5,000}{2,000,000} = .0025$ (or

25 out of every 10,000 students). We could, for example, take a randomly sorted list of the 2,000,000 students, select every 400th name on the list, and we would obtain a sample of 5,000 names; each of the names had a .0025 chance of being selected into the sample. The sampling interval of 400 was obtained by dividing the total population (2,000,000) by the number of desired sample elements (5,000). Note that this sampling interval is the reciprocal of the selection probability: $\frac{1}{.0025} = 400$, or the *sampling weight*. Each person in our hypothetical sample represents 400 people from the original population.

In actuality, we do not have a list of all freshmen in the US, so it is not feasible to simply randomly sample 5,000 of the 2,000,000 students. We might, instead, draw a multistage sample (as discussed in Desideratum 4) by selecting institutions and then sampling students within each selected institution. The difficulty with this approach, however, is that we now need to determine two selection probabilities, one for the institutions (π_j) and one for the freshmen within the selected institutions (π_{ij}). In order to obtain a sample of 5,000 students, the product of these two probabilities must equal the overall desired selection probability of .0025. We can arbitrarily select a value for one of these probabilities and the other value will therefore be determined. For example, suppose we decide to sample 5% (or .05) of the students within each selected college. We would then need to select 5% of the colleges because $.05 \times .05 = .0025$. Sampling 5% of the colleges ($\pi_j = .05$) and sampling 5% of the freshmen at each selected college ($\pi_{ij} = .05$), we would obtain a sample with an expected size of 5,000 and each element in the population would have an overall selection rate of .0025 ($\pi_{ij} = \pi_j \times \pi_{i|j} = .05 \times .05 = .0025$). But note that the number of freshmen typically varies across colleges and with this proposed process we might sample a relatively large number of students in very large colleges (for example, with a sampling rate of 5% and 5,000 freshmen in a college we would have a sample of 250 freshmen at that college) and we might sample only 1 freshman at another college (because there might be only 20 freshmen in total at the college). For the sake of efficiency, it is more typical to conduct surveys in a standardized manner across PSUs and with this sampling plan of using a fixed sampling rate within all institutions we could not guarantee a specific size of the sample at each college. An alternate plan might be to sample institutions at a fixed selection rate and then to sample a specified number of freshmen at each college, for example, 20 students. With this sampling design of taking a specified number of participants at each site, in a two-stage sample, the students in small schools have a very high probability of selection into the sample if their college is selected (given the example numbers above, the conditional selection rate would be $\pi_{i|j} = \frac{20}{20} = 1.00$).

Conversely, students who are in very large colleges have a relatively small chance of being selected for the sample if their college is chosen (for example, $\pi_{ij} = \frac{20}{5000} = .004$). If the colleges are sam-

pled with equal probabilities, then students from these two different colleges would have very different overall rates of selection: $\pi_{ij} = \pi_j \times \pi_{ij} = .05 \times 1.00 = .05$ for students in small colleges and $\pi_{ij} = \pi_j \times \pi_{ij} = .05 \times .004 = .0002$ for students in large colleges. Therefore, freshmen in small schools would be over-represented in the sample, selected at a rate 250 times that of the students in the larger colleges. This overrepresentation can be handled in analyses by using sampling weights.

Students in our example small college would have a weight of $w_s = \frac{1}{.05} = 20$ and for students in our larger college, $w_l = \frac{1}{.0025} = 500$.

A way to avoid having unequal numbers of selected students per colleges or vastly unequal overall probabilities of selection for individual students across colleges is to sample colleges not with equal probability but rather to select them with a method that samples larger institutions at higher rates and smaller institutions at lower rates. This method, called *probability proportionate to size* (PPS) *sampling*, was introduced in Desideratum 4 and is commonly used in national data collection efforts. With PPS sampling, the overall selection probabilities for sample members (π_{ij}) will tend to be similar, but their conditional probabilities (π_{ij}) within their respective PSUs will differ, as will the selection probability for their PSU. For example, suppose with our previous example that we want to select about 20 freshmen at each college no matter the size. A college with 20 freshmen would need to have a .0025 chance of selection into the sample (that is, $\pi_j = .0025$), to result in an overall probability of inclusion for a student in that college of .0025 ($\pi_{ij} = \pi_j \times \pi_{ij} = .0025 \times 1.0 = .0025$). Alternately, a college with 5,000 freshmen would need to have a chance of selection of .625 in the sample ($\pi_j = .625$), to result in an overall selection probability for the students in those colleges of .0025 ($\pi_{ij} = \pi_j \times \pi_{ij} = .625 \times .004 = .0025$). The use of PPS is one source of obtaining sample members with approximately equal sampling weights but this equivalency is not guaranteed.

Disproportionate sampling rates across strata may also be used, and are employed to obtain sufficient numbers of elements to undertake subgroup reporting. When the desire is to report estimates by subgroup, sample designers might employ a higher rate of sampling in certain strata than the rate used in other strata. For example, continuing our example, special interest might lie in reporting estimates for African American freshmen and, therefore, instead of selecting 20 freshmen at random at each institution, researchers might select 5 African American freshmen and 15 freshmen of other race/ethnic groups. This method of selection will likely result in conditional sampling rates for African American students that are greater than that of other students within the same institution (for example, if there are 100 African American freshmen and 900 non-African American freshmen at the college, then the conditional sampling rate for African American students is $\pi_{ij} = \frac{5}{100} = .05$ and for other students in the same institution the sampling rate would be $\pi_{ij} = \frac{15}{900} = .017$. These differing conditional probabilities will lead to different sampling weights.

Finally, as suggested in Desideratum 7, post-stratification weighting adjustments might be employed by the survey designers to adjust for non-response. Although equal selection probabilities might have been used for all elements in the initial sampling plan, some groups typically respond at lower rates than others. After survey data are collected, sampling weights for responses from underrepresented groups (given their response rate) would be set higher to reflect their true proportion in the population (if it is known or can be approximated) and the sampling weights for proportionately over-represented groups would be adjusted lower. It is important to note that the

use of post-stratification weighting for non-response can be somewhat controversial (Lohr, 1999) and several methods exist to adjust for the non-response including cell weighting, raking, regression estimation, and more complex modeling approaches.

Statistically, the use of sampling weights to address non-response reduces bias in parameter estimates but it should be noted that bias is reduced at the expense of precision. If the non-response in fact is not related to the response variable of interest, the parameter estimate will not be biased and weights to adjust for disproportional non-response will not be needed. However the use of weights (either developed through post-stratification adjustment or due to initial disproportionate selection probabilities) in this situation will result in estimated standard errors that will be larger than if they had been estimated as if from an equal probability of selection sample. Therefore, the weighted analysis becomes less powerful. Authors should be explicit about the sampling weights available for their data and whether and how they were included in the analysis. Some authors may opt to analyze the data both weighted and unweighted to examine the effects of the inclusion of weights and provide both pieces of information in their manuscript.

9. Design Effects

A *design effect* refers to an inflation or deflation in the sampling variance (or square of the standard error) of a statistic due to the chosen sampling design. Under an assumption of simple random sampling (SRS), the sampling variance (sv) for the estimate of a population mean is typically defined in textbooks and in software packages as

$$sv(\hat{\mu}) = \frac{s_y^2}{n} \quad (1)$$

where

$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \hat{\mu})^2}{(n-1)} \quad (2)$$

Thus, as sample size increases, the estimate of the sampling variance of $\hat{\mu}$ decreases and the estimate $\hat{\mu}$ of becomes more *efficient*. Technically, Equation (1) assumes that observations were sampled with replacement, which is not typically the case in practice. However, when sampling has been undertaken without replacement, the use of Equation (1) to obtain an estimate of the sampling variance can be acceptable if the sampling fraction is small. The procedure of determining the precision of a parameter estimate is referred to as *variance estimation*. It is more typical in applied research, however, for authors to refer to the estimate of the standard error (se) for the parameter estimates. This se is just the square root of the sampling variance:

$$se(\hat{\mu}) = \sqrt{sv(\hat{\mu})} = \frac{\sqrt{s_y^2}}{\sqrt{n}} = \frac{s_y}{\sqrt{n}} \quad (3)$$

When analyzing data that have been collected through some sampling designs described in Desideratum 4, one of the main problems is that the estimates of sampling variances (and thus standard errors) will be biased when using the traditional formulas in Equations (1) and (3). Use of these formulas includes the important assumption that observations are independent. When data have been collected using complex sampling designs, the observations often are *not* independent and thus sampling variances calculated via Equation (1) will be biased. A measure of how biased a traditional sampling variance estimate will be is called the *design effect* (DEFF). DEFF is the ratio of the correct

sampling variance of a statistic under the complex sampling design over the sampling variance that would have been obtained had SRS been used (Heeringa et al., 2010; Kish, 1965). The square root of DEFF (termed DEFT) is the estimate of the bias in the standard error estimate. If the complex sampling design has no effect on the sampling variance, the value of the DEFF would be 1.0. If the sample design improves the precision of the parameter estimate (such as with stratification), the DEFF will be less than 1.0, and if the design lessens the precision (as found with multistage sampling), the DEFF will be greater than 1.0. To estimate the DEFF for a sample mean, one must obtain an estimate of the appropriate sampling variance using software that can accommodate the sampling design (such as STATA, SUDAAN, WesVar, and select procedures in SAS) and calculate the “naïve” sampling variance assuming SRS with any typical software. The estimate of DEFF for the sample mean, then, is

$$DEFF = \frac{sv(\hat{\mu})_{complex}}{sv(\hat{\mu})_{SRS}} \quad (4)$$

Authors can provide these DEFFs for key variables in analyses to allow the reader to determine whether statistical procedures to accommodate the sampling designs should be used or whether traditional statistical procedures will yield sufficiently robust standard error estimates.

If software that can accommodate survey design information is not available to the researcher, it is still possible to obtain an estimate of the appropriate sampling variance of the sample mean for select sampling designs. Under multistage sampling with no stratification, an estimate of the DEFF of the sample mean can be obtained based on the *intraclass correlation* (ICC). The ICC is a measure of the amount of variability in the response variable y that can be explained by the fact that sample members were selected from within PSUs as opposed to selected randomly. The ICC typically ranges from 0 to 1 and a value close to 1 indicates that all of the elements in the PSU are nearly identical and therefore most variance is found between the PSU means as opposed to within PSUs. An ICC value near zero indicates that, within PSUs, the individuals differ on the response variable and the PSU means do not differ greatly. An estimate of the ICC for a given sample, $\hat{\rho}$, can be obtained using components from an analysis of variance on the variable of interest using the PSU identifier as the between subjects factor:

$$\hat{\rho} = \frac{MS_B - MS_W}{MS_B + (n_s - 1)MS_W} \quad (5)$$

Where MS_B is the model mean square, MS_W is the mean square error, and n_s is the sample size per group if balanced. The design effect of the mean is thus

$$DEFF = 1 + (n_s - 1)\hat{\rho} \quad (6)$$

Providing DEFF information for key variables in the analysis alerts the reader to information about possible violation (or lack of violation) of independence assumptions.

10. Variance Estimation

Variance estimation refers to the estimation of appropriate sampling variances (or standard errors) for a given sample statistic (such as a mean, regression coefficient, or factor loading). When data are not collected with SRS, typical formulas for sampling variance and standard error estimations are not appropriate.

The first decision that a researcher must make is whether the sampling strategy is to be part of the analytic model or whether the sampling strategy will be accounted for in the analysis. The

former, referred to as a *model-based analysis*, presumes that the sampling information is helpful in explaining the hypothesized relations within the data. For example, if a two-stage survey of students within universities was conducted and students were asked about alcohol consumption and location of residence (on-campus housing, fraternity or sorority, off-campus with friends, off-campus with parents), a researcher might posit that the relation between location of residence and alcohol consumption actually depends on the type of university the student is attending, referred to as a cross-level moderation. In that case, the researcher might want to include university information (sampling information) into the analytic model. Techniques such as hierarchical linear modeling (see Chapter 22, this volume) allow the researcher to model the nested structure of the data resulting from the two-stage sampling design.

For those researchers who do not hypothesize that the sampling information is part of the analytic model, that information can be used to appropriately estimate the sampling variances (or standard errors) for the traditional single-level analysis. Such analyses are referred to as *design-based analyses*. The remainder of this section will refer to sampling variance estimation in design-based analyses.

When a sample has been collected via multistage sampling, the estimates of parameters (such as regression coefficients, means, or correlations) typically will be less precise than had a sample of the same size been collected through SRS. Therefore, if a researcher analyzes the data assuming independence of observations, the standard errors for estimates and thus the confidence intervals around the estimates will be underestimated or too narrow; the degree of this bias will depend on the homogeneity of the response variable(s) across clusters as can be measured by the design effect (see Desideratum 9).

If stratification is used as part of the sampling design and the response variable is homogenous within strata, the estimates from the sample will be more precise than had a sample of the same size been obtained through SRS and thus DEFF is less than 1.0. Therefore, if a researcher analyzes the data ignoring the fact that stratification was used in the sampling design, standard errors associated with parameter estimates will be overestimated and the researcher will lose power and increase the likelihood of making Type II errors. If a sampling design includes both stratification and multistage sampling, the increase in precision of estimates resulting from stratification tends to be smaller than the decrease in precision found with multistage sampling.

Several methods are available to estimate sampling variances, standard errors, and test statistics when analyzing data collected through complex sampling designs. Design effect adjustments, linearization, and replication techniques are some of these methods, and while some of these methods can be estimated by hand, most will depend on having access to software that can accommodate the sampling information.

A simple method to adjust for a complex sampling design is to inflate the standard errors obtained from a conventional weighted analysis by the DEFT of the mean of the dependent variable(s) in the analysis. Equivalently, a design effect adjusted sampling weight could be calculated and used in the analysis. The procedures of adjusting a standard error estimate by the DEFT or using an adjusted sampling weight result in accurate adjustments of the standard error for a simple statistic such as the mean. However, these procedures often result in conservative estimates of the sampling errors in more complex statistical procedures such as regression and structural equation modeling. Additionally, the researcher needs very good estimates of DEFF, which would require software that can accommodate the sampling design information. Therefore, it is suggested that researchers plan to use one of the more advanced techniques as described below.

Estimating sampling variances using a linearization method is a more appropriate approach. Because complex sample statistics are actually nonlinear functions, their sampling variances are often obtained by creating an approximate linear function, and then the variance of the new function is used as the sampling variance estimate. This approach to variance estimation is referred to

by several additional terms in statistical analysis literature: the *delta method*, *Taylor Series approximation*, and *propagation of variance*. In the specific case of complex sample data derived from a stratified multistage sample, linearization results in a variance estimate that is a combination of the variation among PSUs within the same stratum. For example, for a stratified multi-stage sample with equal sample sizes within each PSU in a stratum, the standard error of the mean would be estimated in two steps. First, the sampling variance within each stratum would be estimated

$$s_h^2 = \frac{\sum_{\alpha=1}^a (\bar{y}_{h\alpha} - \bar{y}_h)^2}{a-1} \quad (7)$$

where α represents the PSU, a is the total number of PSUs within the stratum, $\bar{y}_{h\alpha}$ is the mean on the response variable in PSU α within stratum h , and \bar{y}_h is mean of the response variable within stratum h across all PSUs. Then these estimates of sampling variance within each stratum are combined to obtain the overall standard error

$$se_{\hat{\mu}} = \sqrt{\sum_{h=1}^H \frac{s_h^2}{n}} \quad (8)$$

where h represents the stratum, H is the total number of strata, n is the total sample size, and s_h^2 is the variance of PSU means within stratum h as calculated in Equation (7).

There are many options to determine an approximate linear estimate and the choice of these depend on the complexity of the sampling design and the complexity of the parameter being estimated. Equations for linearized estimates for sampling variances for a range of different sampling schemes are available in Kalton (1983). Most researchers, however, use computer software that has been specially designed for complex sample data to provide these linearized estimates; information about software options is provided at the end of this section.

Another option for appropriate estimation of standard errors when using complex sample data is to use a replication method. The phrase “replication method” means that repeated samples are taken of the elements in the original sample to constitute new samples. For each of these new samples, the statistic of interest (a mean, a regression coefficient, etc.) is calculated. Then, the empirical distribution of those statistics is used to determine the estimated sampling distribution. With complex sample data, researchers most often use the following replication methods: *jackknife repeated replication*, *balanced repeated replication*, and *bootstrapping*. A very nice description of each of these methods is available in Rust and Rao (1996). The choice of any of these methods can depend on the statistical software available, whether replicate weights are already provided with the data, and the nature of the complex sample design.

Jackknife repeated replication (JRR) entails temporarily dropping one or more observations from the original dataset, obtaining the estimated statistic based on this subsample, and repeating this process until each observation has been dropped once. With multistage data, this process is usually accomplished by dropping all of the elements in one PSU at a time: the first-stage sampling unit, the PSU, is seen as the “dropped” observation. Sampling weights for elements from the other PSUs in the same strata are then adjusted to account for the dropped observation(s) yielding a sum of weights that is the same as the original sum of weights.

The standard error of the statistic is then calculated as a function of the variability of replicate estimates from the original estimate, although the calculation depends on whether a stratified sample was taken at the first-stage of selection. On some national and international data sets that are publicly available, the adjusted weights are already provided for the analyst. For example, a dataset might include 90 jackknife weights, called JACK1, JACK2, . . . , JACK90, indicating that the analyst

might choose a replication method for standard error estimation, running the analysis 90 times, each time with a different weight. Those 90 estimates will then be used to determine the sampling variability of the original full sample estimates.

Balanced repeated replication (BRR), also referred to as *half-sample replicates*, is an approach where each replicate is created using half of the PSUs in the sample, one from each stratum. A second replicate, the complement replicate, can then be created out of the remaining PSUs. BRR can only be accomplished when the sampling design has been undertaken with the selection of two PSUs from each stratum. If the sample design did not include two PSUs from each stratum, similar strata and/or PSUs can be grouped to obtain such a design but such realignment must be done with caution. The term “balanced” in the name BRR refers to the need to choose orthogonal replicates. There is a complication creating replicates using half of the PSUs (chosen at random) because dependent replicates can result, providing estimated statistics that are correlated across replicates. For example, if we have four strata in our sampling design (strata 1 – 4) and within each stratum sampled two universities (with IDs of 1 to 8 continuous across strata), we could obtain the following replicates with selected university IDs:

Replicate 1: 1, 3, 5, 7

Replicate 2: 1, 4, 5, 8

Replicate 3: 2, 4, 5, 8

Because replicates 2 and 3 share three of the same universities in the replicate sample, the estimated statistic in these two replicates will be very similar. For this reason, design matrices are used to select the appropriate PSUs for each BRR replicate sample; they are not chosen at random). The standard error of the statistic is then determined based on the variability of the estimate across replicates.

Bootstrapping is similar to JRR and BRR in that the observations from the original sample are used to form replicate samples. In bootstrapping, however, observations from the original sample are sampled with replacement to obtain a dataset that can be of the same size as the original dataset and no reweighting of the observations is typically undertaken. Bootstrapping is not an easy task with complex sample data but is a very flexible method. The process of creating bootstrapped replicates will depend on the complex sampling design. The standard error of the estimate will be a function of the variability of the estimate across the many bootstrap replicate samples. Bootstrap methods typically require many more replications than JRR or BRR. Additionally, while JRR and BRR estimation are available in many survey software packages already, bootstrapping must usually be programmed by the researcher.

There are several software options for the researcher who would like to use linearization or replication techniques, in conjunction with their sampling design information, to appropriately estimate sampling variances of statistics. Starting with version 8, SAS has included linearized sampling variance estimates for select procedures (such as means, regression coefficients, and frequency tables.) The WesVar software was developed specifically for analyzing complex sample data and relies on replication techniques to variance estimation; both BRR and JRR are accommodated in this software. SUDAAN supports both JRR and BRR replication methods, as well as linearized variance estimates. Stata, like SAS, is a full data base management and statistical package that also includes a complex sample modeling component, relying on linearization for variance estimation. In SPSS, a module called COMPLEX allows researchers access to advanced functions to specify the sampling design. Finally, specialized statistical software packages, such as LISREL and Mplus for structural equation modeling, now include estimation techniques for data that arise from complex sampling designs.

If authors have used data from a complex sampling design, at a minimum they should alert readers to the violation of assumptions for traditional statistical analyses and the likely effects on standard errors and inference that these violations would present. With the many software programs that are available to appropriately analyze complex sample data, authors should attempt to accommodate the sampling design using one of the sampling variance estimation techniques identified above.

11. Limitations in Survey Data Collection

Researchers are encouraged to remark about anything in the survey data collection and analysis that was not in line with the guidelines above. The context of the survey, however, is the most important consideration and decisions pertinent to this context can easily override the desiderata and other guidelines presented. Remarks about the likely effects of each of the problems or issues that arise for each desideratum would be appropriate, including coverage error in the discussion of the sample frame, measurement error in the discussion of questionnaire development and measures used, non-response error in the discussion of the procedures and response rate, editing and process errors, and, of course, sampling error in the statistical analysis portion.

References

- American Association for Public Opinion Research. (2006). *Standard definitions: Final dispositions of case codes and outcome rates for surveys*. Retrieved January 9, 2008 from www.aapor.org/uploads/standarddefs_4.pdf.
- American Association for Public Opinion Research. (2008). *Best practices for survey and public opinion research*. Retrieved January 10, 2008 from www.aapor.org/bestpractices.
- De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (Eds.) (2008). *International handbook of survey methodology*. New York: Lawrence Erlbaum Associates.
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. New York: John Wiley & Sons.
- Forsyth, B. H., & Lessler, J. T. (1991). Cognitive laboratory methods: A taxonomy. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman, S. (Eds.), *Measurement errors in surveys* (pp. 393–419). New York: Wiley.
- Frank, O., & Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53–67.
- Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.
- Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.
- Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis*. Boca Raton, FL: CRC Press.
- Kalton, G. (1983). *Introduction to survey sampling*. Newbury Park: Sage.
- Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons.
- Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.
- Lohr, S. L. (2008). Coverage and sampling. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 97–112). New York: Lawrence Erlbaum Associates.
- Rust, K. F., & Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283–310.
- Smith, T. W. (2002) Developing nonresponse standards. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 27–40). New York: Wiley.
- Stapleton, L. M. (2008). Analysis of data from complex surveys. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 342–369). New York: Lawrence Erlbaum Associates.
- Stapleton, L. M. (2013). Incorporating sampling weights into single- and multi-level models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 353–388). London: Chapman Hall/CRC Press.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York: Cambridge University Press.

Contributors

Jill L. Adelson (jadelson@tip.duke.edu) is a Research Scientist with the Duke University Talent Identification Program. She has a joint doctorate in Measurement, Evaluation, and Assessment and in Gifted Education from the University of Connecticut. As a translational quantitative methodologist, her research involves bridging the gap between advanced methods and applied research through (a) applications of advanced methods in gifted education, (b) collaborations involving applications of advanced methods, and (c) methodological dissemination. She is the co-editor of *Gifted Child Quarterly* and was the Specialty Chief co-Editor of *Quantitative Psychology and Measurement for Frontiers in Psychology* and *Frontiers in Applied Mathematics and Statistics*.

Paul D. Allison (allison@statisticalhorizons.com) is President of Statistical Horizons and Professor Emeritus of the University of Pennsylvania. After completing his doctorate in sociology at the University of Wisconsin, he did postdoctoral study in statistics at the University of Chicago and the University of Pennsylvania. He has published eight books and more than 70 articles on topics that include linear regression, log-linear analysis, logistic regression, structural equation models, inequality measures, missing data, and survival analysis. At present, his main focus is on the analysis of panel data and on methods for handling missing data. A former Guggenheim Fellow, he received the 2001 Lazarsfeld Award for distinguished contributions to sociological methodology. In 2010 he was named a Fellow of the American Statistical Association.

K. Rivet Amico (ramico@umich.edu) is Associate Professor of Health Behavior and Health Education in the School of Public Health at the University of Michigan. She is an active contributor in the areas of HIV-prevention and treatment social-behavioral theory development, intervention implementation, and evaluation and measurement. Her research includes work with engagement in HIV care, TB and drug resistant TB prevention and treatment, PreExposure Prophylaxis (PrEP), social behavioral factors influencing participants in clinical trials, and measures development, with the overarching goal of advancing the reach and quality of care and prevention services available domestically and internationally. She also has strong interests in research design, implementation science, program evaluation and community capacity building, and her work has appeared in *JAIDS*, *AIDS and Behavior*, and *American Journal of Public Health*.

Deborah L. Bandolos (bandaldj@jmu.edu) is Professor and Director of the Assessment and Measurement Program at James Madison University. Her research areas include structural

equation modeling, exploratory factor analysis, instrument development and validation, and educational accountability and assessment systems. Her research has appeared in such journals as *Structural Equation Modeling: A Multidisciplinary Journal*, *Multivariate Behavioral Research*, *Applied Measurement in Education*, *Educational Measurement: Issues and Practice*, *Psychological Assessment*, *Educational and Psychological Measurement*, and *Journal of Educational Measurement*, and she is the author of *Measurement Theory and Applications for the Social Sciences* (Guilford Publications). She sits on a number of editorial boards of leading methodological journals, including as Associate Editor of *Multivariate Behavioral Research*; she is also an elected member of the Society of Multivariate Experimental Psychology, and a Fellow of the American Psychological Association.

Christine H. Barthold (choffner@gmu.edu) is Assistant Professor at George Mason University, and Coordinator of the Applied Behavior Analysis (ABA) program. Her research experience includes the oversight of two federally funded research projects dedicated to understanding how children with autism use symbols to communicate. She teaches and mentors students learning Single-Subject Design, and often uses Single-Subject designs in her scholarship. Her research interests include the application of ABA to interventions for individuals with and without disabilities as well as using interteaching to increase active student responding in higher education.

S. Natasha Beretvas (tberetvas@austin.utexas.edu) is the John L. and Elizabeth G. Hill Centennial Professor in the Quantitative Methods program in the Department of Educational Psychology and Associate Dean for Research and Graduate Studies in the College of Education at the University of Texas at Austin. She conducts research on meta-analysis both for single-case and group-comparison experimental designs' data and has a research program focused on extensions to the conventional multilevel model for handling student mobility. Her research has appeared in journals such as *Multivariate Behavioral Research*, *Psychological Methods*, and *Behavioral Research Methods*. She is a member of the Society for Research Synthesis Methodology and is editor-in-chief of the *Research Synthesis Methods* journal.

Wanchen Chang (wanchen.chang@nwea.org) is Research Scientist at NWEA, a Pre-K-12 not-for-profit organization in Portland, OR. Her areas of interest include multilevel modeling and structural equation modeling, with applications to educational policy and student assessment. Prior to NWEA, she was Assistant Professor at Boise State University, where she taught statistics and research methods courses. Her work has appeared in journals such as *Journal of Experimental Education* and *General Linear Model Journal*. She has also co-authored methodological studies on multilevel modeling and item response theory.

Brittany F. Crawford (bnflan02@louisville.edu) is a doctoral student in the Educational Psychology, Measurement, and Evaluation program at the University of Louisville. Her research interests include gifted education, motivation, and assessment. She is actively involved in multiple committees for the National Council on Measurement in Education and the Measurement and Research Methodology Division of the American Educational Research Association.

Robert A. Cribbie (cribbie@yorku.ca) is Professor in the Quantitative Methods Program within the Department of Psychology at York University. His research interests include equivalence testing, robust ANOVA, multiplicity control, and the measurement of change.

Geoff Cumming (g.cumming@latrobe.edu.au) is Professor Emeritus in Psychology at La Trobe University, and author of two statistics textbooks, including, with Bob Calin-Jageman, *Introduction to The New Statistics: Estimation, Open Science, and Beyond* (see www.thenewstatistics.com). This introductory statistics textbook is the first to integrate Open Science and the new statistics (estimation and meta-analysis) from the start. He has taught statistics for more than 40 years, and his statistics tutorial articles have been downloaded more than 370,000 times. His main research interests are the investigation of statistical understanding, and promotion of Open Science and improved

statistical practices. A Rhodes Scholar, he received his Doctorate degree in experimental psychology from Oxford University.

Patrick J. Curran (curran@unc.edu) is Professor in the Department of Psychology and Neuroscience at the University of North Carolina at Chapel Hill and Director of the L. L. Thurstone Psychometric Laboratory. His quantitative program of research is primarily focused on the development, evaluation, and application of advanced statistical methods designed to evaluate individual stability and change over time. His substantive program of research is focused on risk and protective factors in developmental trajectories of drug and alcohol use and abuse in children and adolescents. He has dedicated a substantial portion of his time to the dissemination of advanced quantitative methodologies through his teaching, workshops, mentoring, and writing of pedagogically oriented books and articles.

Sharon Anderson Dannels (sdannels@gwu.edu) is Associate Professor of Educational Research and Associate Dean for Doctoral Studies at the Graduate School of Education and Human Development, George Washington University. Her most recent work focuses on the leadership development of mid-career women faculty in traditionally male-dominated disciplines/fields within the academy. Her scholarship includes quantitative and qualitative research in addition to program evaluations. Her work has been funded by the Jim Joseph Foundation, the Alfred P. Sloan Foundation, and the US Department of Education, and has included projects on evaluation protocols, educational facilities, interpreting for deaf individuals with dysfluent language, and women and academic leadership.

C. Mitchell Dayton (cdayton@umd.edu) is Professor Emeritus in the Measurement, Statistics and Evaluation program at the University of Maryland. His primary areas of interest are latent class models and other discrete mixture models. His Sage book, *Latent Class Scaling Analysis*, is a widely recognized resource. His research has appeared in such journals as *Psychometrika*, *Journal of the American Statistical Association*, *American Statistician*, *Multivariate Behavioral Research*, *Applied Psychological Measurement*, *Journal of Educational and Behavioral Statistics*, *British Journal of Mathematical and Statistical Psychology*, *Psychological Methods*, *Journal of Modern Applied Statistical Methods*, and *Annals of Translational Medicine*.

R. J. De Ayala (rdeayala2@unl.edu) is Professor and Chair of the Department of Educational Psychology at the University of Nebraska – Lincoln. His research interests in psychometrics, item response theory, and computerized adaptive testing has appeared in *Applied Psychological Measurement*, *Applied Measurement in Education*, *British Journal of Mathematical and Statistical Psychology*, *Educational and Psychological Measurement*, *Journal of Applied Measurement*, *Multivariate Behavioral Research*, and *Journal of Educational Measurement*. He has also authored *The Theory and Practice of Item Response Theory* and serves on the editorial boards of Educational and Psychological Measurement and Applied Measurement in Education. He is a Fellow of the American Psychological Association and the American Educational Research Association.

Cody Ding (dingc@umsl.edu) is Professor at the University of Missouri-St. Louis. His area of interest includes multidimensional scaling, with a particular focus on latent growth and latent mixture profile models. He is also interested in educational and psychological assessment, psychosocial development, social trauma, learning, and the neural processes of learning and behaviors, and his work has appeared in such journals as *International Journal of Behavioral Development*, *International Journal of Psychology, Measurement and Evaluation in Counseling Development*, and *Biological Psychology*. He authored the book *Fundamentals of Applied Multidimensional Scaling in Education and Psychological Research* (Springer).

Andrew L. Egel (aegel@umd.edu) is Professor Emeritus of Counseling, Higher Education and Special Education at the University of Maryland. His area of interest includes the development

and evaluation of innovative instructional methodologies for students with an Autism Spectrum Disorder and training teachers and parents in the implementation of evidence-based strategies. His work has appeared in such journals as *Journal of Applied Behavior Analysis*, *Journal of Autism and Developmental Disorders*, and *Review Journal of Autism and Developmental Disorders*.

Monica K. Erbacher (mkerbacher@email.arizona.edu) is Assistant Professor in the Department of Educational Psychology in the College of Education at the University of Arizona. She teaches a variety of applied statistics courses on topics ranging from basic inferential statistics to multivariate methods (e.g., cluster analysis, factor analysis), advanced linear models (e.g., hierarchical linear modeling), and measurement theories (e.g., item response theory). Her research uses advanced statistical techniques to develop and compare models of student attitudes and behaviors, particularly academic entitlement. Her publications have appeared in the *Journal of Applied Measurement*, *Psychological Assessment*, *Structural Equation Modeling*, and the *College Student Affairs Journal*.

Xitao Fan (xtfan@cuhk.edu.cn) is the Presidential Chair Professor and Dean of the School of Humanities & Social Science, Chinese University of Hong Kong (Shenzhen), China. Previously, he served at the University of Macau, China (Chair Professor, School Dean, and Interim Vice President for Academic Affairs), University of Virginia (Professor, and Endowed Chair), and Utah State University. His interest areas include applied quantitative methods, educational and psychological measurement, multivariate modeling, meta-analytic synthesis, and inter-disciplinary research. His work has appeared in *Multivariate Behavioral Research*, *Psychological Methods*, *Structural Equation Modeling: A Multidisciplinary Journal*, *Educational Psychology Review*, *American Educational Research Journal*, *Educational and Psychological Measurement*, and others.

Fiona Fidler (fidlerfm@unimelb.edu.au) is Associate Professor at the University of Melbourne, where she holds a joint appointment in the School of BioSciences and the School of Historical and Philosophical Studies. She is a current Australian Research Council Future Fellow, investigating questionable research practices and their effects in ecology and evolution. She is interested in how methodological change occurs in science, and more generally, how scientists and other experts reason and make decisions.

Sara J. Finney (finneysj@jmu.edu) is Professor and Associate Director of the Center for Assessment and Research Studies at James Madison University. She teaches courses in multivariate statistics and structural equation modeling, and advises graduate students in the Assessment & Measurement Ph.D. program and the Quantitative Psychology M.A. program. In her role as Associate Director, she works with faculty and staff to design assessment efforts that contribute to making empirically based decisions about program effectiveness and ultimately student learning and development. Her research involves the application of latent variable modeling techniques to investigate questions related to college student development, examinee motivation, and the functioning of self-report instruments, and has appeared in such journals as *Applied Measurement in Education*, *Educational and Psychological Measurement*, *Educational Assessment*, and *Journal of Experimental Education*.

Gregory R. Hancock (ghancock@umd.edu) is Professor, Distinguished Scholar-Teacher, and Director of the Measurement, Statistics and Evaluation program as well as the Center for Integrated Latent Variable Research (CILVR) at the University of Maryland. His area of interest includes structural equation modeling with a particular focus on latent growth and latent means models, and his work has appeared in such journals as *Psychometrika*, *Multivariate Behavioral Research*, *Psychological Methods*, *Structural Equation Modeling: A Multidisciplinary Journal*, *British Journal of Mathematical and Statistical Psychology*, and *Journal of Educational and Behavioral Statistics*. He has co-edited a number of methodological volumes, is an elected member of the Society of Multivariate Experimental Psychology, and is a Fellow of the American Psychological Association, Association for Psychological Science, and the American Educational Research Association.

Amy Hendrickson (ahendrickson@collegeboard.org) is Senior Director, Psychometrics at the College Board. Previously, she was Assistant Professor in the Department of Measurement, Statistics, and Evaluation at the University of Maryland. Her areas of interest are in large-scale operational issues and analyses concerning item and form analyses, including reliability and Generalizability Theory; IRT; test security; and linking and scaling. Her publications have appeared in *Educational Measurement: Issues and Practice*, *Journal of Educational Measurement*, *Applied Measurement in Education*, and *The Handbook on Measurement, Assessment, and Evaluation in Higher Education*.

William T. Hoyt (wthoyt@wisc.edu) is Professor in the Department of Counseling Psychology and Associate Dean of the School of Education at the University of Wisconsin-Madison, and an affiliate faculty member in the Department of Educational Psychology (Quantitative Methods program). As an applied psychologist and research methodologist, he has interests in strengthening the evidentiary basis for practice, including research synthesis (meta-analysis), appropriate interpretations of statistical models, and validity of measurement (including reliability and generalizability analysis).

Paul E. Jose (paul.jose@vuw.ac.nz) is Professor of Psychology in the School of Psychology at Victoria University of Wellington in New Zealand. He teaches and performs research in the fields of developmental psychology and positive psychology, but has a passionate interest in developmental statistics and methodology as well. He served as Associate Editor for journal *Developmental Psychology* for five years, and has performed reviews for many other journals in developmental psychology. In terms of research productivity in methods and statistics, he published the book *Doing Statistical Mediation and Moderation* (Guilford Press, 2013), and the paper “The merits of using longitudinal mediation” (*Educational Psychologist*, 2016).

Ken Kelley (kkelley@nd.edu) is the Edward F. Sorin Society Professor of IT, Analytics, and Operations (ITAO) and the Associate Dean for Faculty and Research in the Mendoza College of Business at the University of Notre Dame. He is in the analytics group within the ITAO Department and works to advance analytic methods. His specialties are in the areas of designing studies and analyzing data, particularly effect size estimation and confidence interval formation, longitudinal data analysis, and statistical computing. In addition to his methodological work, he collaborates with colleagues on a variety of important topics developing new and applying existing but non-standard methods to a variety of areas in business and psychology. He is an Accredited Professional Statistician (PStat) by the American Statistical Association, former Associate Editor of *Psychological Methods*, recipient of the Anne Anastasi early career award by the American Psychological Association’s Division of Evaluation, Measurement, & Statistics, an elected member of the Society of Multivariate Experimental Psychology, and a Fellow of the American Psychological Association.

Harvey Keselman (Harvey.Keselman@umanitoba.ca) is Professor Emeritus in the Department of Psychology, University of Manitoba. His areas of interest include the analysis of repeated measurements, multiple comparison procedures, and measures of effect size. His research has been published in journals such as *Psychological Bulletin*, *British Journal of Mathematical and Statistical Psychology*, *Educational and Psychological Measurement*, and the *Journal of Modern Applied Statistical Methods* (where he also served on the editorial board). His research has received support from the Social Sciences and Humanities Research Council, and the Natural Sciences and Engineering Research Council of Canada.

Se-Kang Kim (sekim@fordham.edu) is Associate Professor in the Department of Psychology at Fordham University and Director of the Ph.D. Program in Psychometrics and Quantitative Psychology. He was Associate Editor of Applied Psychological Measurement and an editorial board member for Psychological Assessment. His current research includes applications of correspondence analysis and profile analysis utilizing multivariate statistics. His work has appeared in such journals as *International Journal of Methods in Psychiatric Research*, *Journal of Classification*, *Methodology*,

British Journal of Mathematical and Statistical Psychology, Behavior Research Methods, Multivariate Behavioral Research, Psychological Assessment, and American Educational Research Journal.

Alan J. Klockars (klockars@u.washington.edu) is Professor Emeritus of Education at the University of Washington, where he taught for 42 years in the area of Measurement & Statistics. A major area of interest concerns issues in the application of experimental designs and specifically ANOVA designs. These include multiple comparisons and analysis of trait by treatment interactions. His work has appeared in such journals as *Psychological Bulletin*, *Psychological Methods*, *Journal of Educational and Psychological Measurement*, *Journal of Modern Applied Statistical Methods* (where he was Assistant Editor), *Journal of Educational Measurement*, *British Journal of Mathematical and Statistical Psychology*, and *Educational and Psychological Measurement*.

Thomas R. Knapp (tomknapp5@gmail.com) is Professor Emeritus of Education and Nursing, University of Rochester and The Ohio State University. His specialty is the reliability of measuring instruments. He has published five books and about 100 articles in peer-reviewed journals such as *American Educational Research Journal*, *Educational and Psychological Measurement*, *Research in Nursing and Health*, and *Clinical Nursing Research*. He served for many years as a reviewer for research journals in education and nursing.

Timothy R. Konold (tk2e@virginia.edu) is Professor and Director of the Research, Statistics, and Evaluation program in the Curry School at the University of Virginia, and holds faculty affiliations with the Center for the Advanced Study of Teaching and Learning (CASTL), the Virginia Education Science Training (VEST) program, and the Youth Violence Project (YVP). He has authored more than 100 peer-reviewed articles, book chapters, and published tests. His research interests are in structural equation modeling. Examples of journals in which his work has appeared include *Structural Equation Modeling: A Multidisciplinary Journal*, *Educational and Psychological Measurement*, *Psychological Assessment*, and the *Journal of Experimental Education*.

Jennifer L. Kouo (jkouo@towson.edu) is Assistant Professor within the Department of Special Education at Towson University. Her area of interest includes single-case design, with a particular emphasis on its use to study interventions to support individuals with an autism spectrum disorder. Her work has appeared in such journals as *Focus on Autism and Other Developmental Disabilities* and *Review Journal of Autism and Developmental Disorders*.

Stephanie T. Lane (slane@ida.org) is a Research Staff Member at the Institute for Defense Analyses in Alexandria, VA. Her research spans latent variable models of change and variable selection, with emphasis on application to longitudinal and time series data. She is the author of multiple open source software packages that aim to disseminate novel statistical methods to the broader research community. Her research has appeared in outlets such as *Psychological Methods*, *Structural Equation Modeling: A Multidisciplinary Journal*, *Multivariate Behavioral Research*, *Biological Psychiatry*, and the *International Journal of Psychophysiology*.

Lisa Lix (Lisa.Lix@umanitoba.ca) is Professor in the Department of Community Health Sciences, University of Manitoba, and Tier I Canada Research Chair in Methods for Electronic Health Data Quality. She is also Director of the Data Science Platform in the George & Fay Yee Centre for Healthcare Innovation, University of Manitoba. Her areas of methodological research include the analysis of longitudinal/repeated measures data, robust statistical inference, and the analysis of patient-reported outcomes. Her research is supported by the Canadian Institutes of Health Research and the Natural Sciences and Engineering Research Council of Canada. Her research has appeared in *Quality of Life Research*, *BMC Health Services Research*, *Journal of Clinical Epidemiology*, as well as numerous clinical journals. She is a long-standing member of the Statistical Society of Canada and collaborates extensively with colleagues from the Public Health Agency of Canada.

Richard G. Lomax (lomax.24@osu.edu) is Professor Emeritus of Educational and Human Ecology at the Ohio State University, where he taught courses in structural equation modeling and univariate and multivariate statistics, and is former Associate Dean for Research and Administration. His research primarily focuses on early literacy and statistics. He has published in such diverse journals as *Reading Research Quarterly*, *Child Development*, *Journal of Speech, Language, and Hearing Research*, *Journal of Educational Psychology*, *Journal of Negro Education*, *Violence Against Women*, *The School Community Journal*, *Journal of Early Adolescence*, *Parenting: Science and Practice*, *Journal of Counseling and Development*, *American Statistician*, *Journal of Reading Recovery*, and *Journal of Experimental Education*. He is a Fellow of the American Educational Research Association, has 11 published statistics textbooks, and is a three-time Fulbright Scholar.

Gitta Lubke (glubke@nd.edu) is Professor at the University of Notre Dame in the Department of Psychology where she is a member of the Quantitative Area Program. Her general research interests are in the areas of latent variable modeling and data mining. More specifically, she is interested in adapting current data mining approaches to social sciences data which, in addition to longitudinal or hierarchical structures, are often characterized by a rather low signal to noise ratio, and in translating knowledge obtained from data mining to more confirmatory structural equation models. Current substantive work aims at building a predictive model of childhood aggression by combining data mining approaches and integrative data analysis applied to large multi-country data collections.

Fayez S. Maajeeny (fmaajeeny@uj.edu.sa) is Assistant Professor of Special Education at the University of Jeddah. His area of interest focuses on instructional methods of teaching students with autism spectrum disorders, strategies for inclusion of students with disabilities, empirically based instructional methodology, applied behavior analysis, and single-case research methodology. He has a numerous conference presentations on the use of various technologies with students who have autism spectrum disorders.

Scott E. Maxwell (Scott.E.Maxwell.1@nd.edu) is Professor Emeritus at the University of Notre Dame, where he was previously the Fitzsimons Professor of Psychology. His research interests are in the areas of research methodology and applied behavioral statistics, with much of his recent work focusing on statistical power and accuracy in parameter estimation, especially in randomized designs. He has served as editor of *Psychological Methods*; received the Samuel J. Messick Award for Distinguished Scientific Contributions by the American Psychological Association's Division of Evaluation, Measurement, and Statistics; is an elected member of the Society of Multivariate Experimental Psychology, a Fellow of the American Psychological Association, and a Fellow of the Association of Psychological Science; and has received multiple teaching awards.

D. Betsy McCoach (betsy.mcccoach@uconn.edu) is Professor in the Research Methods, Measurement, and Evaluation program in the Neag School of Education at the University of Connecticut. She has co-authored over 100 peer-reviewed journal articles, book chapters, and books, including *Instrument Design in the Affective Domain* and *Multilevel Modeling of Educational Data*. She founded the Modern Modeling Methods conference, held annually at the University of Connecticut. She is also the Director of DATIC, which hosts workshops on a variety of modeling methods. She is co-Principal Investigator for the National Center for Research on Gifted Education and has served as Principal Investigator, co-Principal Investigator, and/or research methodologist for several other federally funded research projects/grants. Her research interests include multilevel modeling, longitudinal modeling, instrument design, latent variable modeling, and gifted education.

Daniel McNeish (dmcneish@asu.edu) is Assistant Professor in the Quantitative Area of the Psychology Department at Arizona State University. His research interests include structural equation modeling, longitudinal data analysis, and modeling of clustered data, particularly with

challenging data structures such as small samples or missing values. His research has appeared in quantitative and substantive journals such as *Psychological Methods*, *Multivariate Behavioral Research*, *Structural Equation Modeling: A Multidisciplinary Journal*, *Educational Researcher*, and *Educational Psychologist*. His work has also been acknowledged with the APA Division 5 Anne Anastasi Dissertation Award, and he is an elected member of the Society for Multivariate Experimental Psychology.

Ralph O. Mueller (rmueller@pnw.edu) is Vice Chancellor for Academic Affairs and Provost at Purdue University Northwest. Previously, he served as Dean of the College of Education, Nursing and Health Professions at the University of Hartford and as Department Chair and Professor at The George Washington University. His writings include textbooks, chapters, and articles on proper applications of multivariate statistical techniques, especially structural equation modeling. He served as editorial board member of *Educational and Psychological Measurement* and *Measurement and Evaluation in Counseling and Development* for more than a decade and was a charter member of the board of *Structural Equation Modeling: A Multidisciplinary Journal*.

Kevin R. Murphy (krm10@me.com) holds the Kemmy Chair of Work and Employment Studies at the University of Limerick. He is the author of over 190 articles and book chapters, and author or editor of 11 books, in areas ranging from psychometrics and statistical analysis to individual differences, performance assessment, and honesty in the workplace. He has served as President of the Society for Industrial and Organizational Psychology and Editor of *Journal of Applied Psychology* and of *Industrial and Organizational Psychology: Perspectives on Science and Practice*. He has served as Chair of the U.S. Department of Defense Advisory Committee on Military Personnel Testing, and has also served on five U.S. National Academy of Sciences committees, all of which dealt with problems in the workplace.

Ann A. O'Connell (oconnell.87@osu.edu) is Professor of Quantitative Research, Evaluation and Measurement and Director of the Research Methodology Center at the College of Education and Human Ecology at The Ohio State University. Her areas of specialization include generalized linear and mixed models, and methods for assessing health and education interventions. Among her published works is a book with Sage on *Logistic Regression Models for Ordinal Response Variables*, and a co-edited volume on *Multilevel Modeling of Educational Data* (with D. Betsy McCoach). She is a former Fulbright Scholar to Addis Ababa University (AAU) in Ethiopia from 2013–2014. She is dedicated to efforts that improve capacity for research methodology locally and abroad.

Jason W. Osborne (jwo@clemson.edu) is Associate Provost for Graduate Studies, Dean of the Graduate School, and Professor of Applied Statistics at Clemson University. His interests in applied statistics have centered around evidence-based best practices spanning topics throughout the behavioral, social, and health sciences. He is author of seven books and numerous peer-reviewed articles; his corpus of scholarship has been cited over 17,000 times.

Dena A. Pastor (pastorda@jmu.edu) has a dual appointment at James Madison University as Professor in the Department of Graduate Psychology and as Associate Director of Assessment Operations in the Center for Assessment and Research Studies. She teaches courses in hierarchical linear modeling, categorical data analysis, and data management. Her research applies statistical and psychometric techniques to the modeling and measurement of college student learning and development. Her publications have appeared in *Contemporary Educational Psychology*, *Applied Psychological Measurement*, and *Applied Measurement in Education*.

Keenan A. Pituch (kpituch@austin.utexas.edu) is Associate Professor in the quantitative methods program in the Educational Psychology Department at the University of Texas at Austin. His areas of interest include multivariate methods, multilevel modeling, mediation analysis, missing data

analysis, and intensive longitudinal modeling. His work has appeared in such journals as *Sociological Methods and Research*, *Multivariate Behavioral Research*, *American Journal of Evaluation*, *Journal of Experimental Education*, and *Methodology*. He is co-author, along with James P. Stevens, of *Applied Multivariate Statistics for the Social Sciences: Analyses with SAS and IBM's SPSS* (6th ed.).

Kristopher J. Preacher (kris.preacher@vanderbilt.edu) is Professor of Quantitative Methods at Vanderbilt University. His research concerns the use of structural equation modeling and multilevel modeling to analyze longitudinal and correlational data. Other interests include developing techniques to test mediation and moderation hypotheses, bridging the gap between theory and practice, and studying model evaluation and model selection in the application of multivariate methods to social science questions. He coauthored the book *Latent Growth Curve Modeling*, serves as Associate Editor of *Psychological Methods*, and serves on the editorial boards of *Behavior Research Methods*, *Communication Methods and Measures*, *Journal of Educational and Behavioral Statistics*, and *Multivariate Behavioral Research*.

David Rindskopf (drindskopf@gc.cuny.edu) is Distinguished Professor of Educational Psychology and Psychology at the CUNY Graduate Center. His areas of research include categorical data, latent variable models (factor analysis, latent class analysis, structural equation models), Bayesian statistics, analysis of single case designs, meta-analysis, and multilevel models. He is a Fellow of the American Statistical Association and the American Educational Research Association. He is past editor of *Journal of Educational and Behavioral Statistics*, and past President of the Society of Multivariate Experimental Psychology.

Karen Samuelson (ksamuelson@piedmont.edu) is Professor, and Director of Doctoral Research in the School of Education at Piedmont College in Athens, Georgia. She co-edited *Advances in Latent Variable Mixture Models* (with Gregory Hancock), and has written chapters in several books on the topic of mixture models. She also has journal articles and book chapters on the topic of validity. Her current area of interest is in the preparation of teachers especially as it pertains to the structure of their college coursework.

Michael A. Seaman (mseaman@sc.edu) is Associate Professor of Educational Research at the University of South Carolina. His primary area of interest is nonparametric statistical methods for behavioral and social research. His work has appeared in journals such as *Psychological Methods*, *Journal of Experimental Education*, *Journal of Educational Measurement*, *Communications in Statistics*, and *Educational and Psychological Measurement*. He is actively involved in funded research projects, with current and past funding from the National Institute of Justice, The National Science Foundation, The National Oceanic and Atmospheric Administration, and the US Department of Education.

Ronald C. Serlin (rcserlin@wisc.edu) is Professor Emeritus, Distinguished Teaching Award winner, former Chair of the Department of Educational Psychology at the University of Wisconsin-Madison, and former Editor of the American Educational Research Association Special Interest Group—Educational Statisticians Book Series. His areas of interest include nonparametric and multivariate statistical methods and the philosophy of science and statistics, and his work has appeared in such journals as *American Psychologist*, *Psychological Methods*, *British Journal of Mathematical and Statistical Psychology*, and *Journal of Educational and Behavioral Statistics*. He co-authored, with Leonard Marascuilo, the text *Statistical Methods for the Social and Behavioral Sciences*.

Laura M. Stapleton (Lstaplet@umd.edu) is Professor in the Measurement, Statistics and Evaluation program in the Department of Human Development and Quantitative Methodology at the University of Maryland. Additionally, she serves as the Associate Director of the Research Branch of the Maryland State Longitudinal Data System Center. Her area of interest is analysis of administrative

data and survey data obtained under complex sampling designs, multilevel latent variable models, tests of mediation within a multilevel framework. Her work appears in such journals as *Journal of Educational and Behavioral Statistics*, *Structural Equation Modeling: A Multidisciplinary Journal*, and *Multivariate Behavioral Research*.

Elizabeth A. Stuart (estuart@jhu.edu) is Professor in the Departments of Mental Health, Biostatistics, and Health Policy and Management at the Johns Hopkins Bloomberg School of Public Health, and Associate Dean for Education at JHSPH. She is a Fellow of the American Statistical Association and has received the mid-career award from the Health Policy Statistics Section of the American Statistical Association and the Gertrude Cox Award for applied statistics. Dr. Stuart has published influential papers on propensity scores and methods to assess the generalizability of randomized trials.

Tracy M. Sweet (tsweet@umd.edu) is Assistant Professor in the Measurement, Statistics and Evaluation program at the University of Maryland. Her research focuses on social network models with particular focus on models used in educational applications. Her methodological work can be found in *Social Networks* and *Journal of Educational and Behavioral Statistics*, as well as in the *Handbook of Mixed Membership and Their Applications*. Applications of her network models are published in *American Journal of Education Research*, *Sociology of Education*, and *American Journal of Education*.

Keith F. Widaman (keith.widaman@ucr.edu) is Distinguished Professor in the Graduate School of Education at the University of California at Riverside. His quantitative areas of interest include exploratory and confirmatory factor analysis, structural equation modeling, and longitudinal modeling, and his substantive interests include growth and development of mental abilities, adaptive behaviors of persons with intellectual disability, and adolescent and young adult development in cultural context. He is a past Editor and current Associate Editor of *Multivariate Behavioral Research*, and his work has appeared in *Psychological Methods*, *Multivariate Behavioral Research*, *Structural Equation Modeling: A Multidisciplinary Journal*, and *Psychometrika*, among others. He is a past winner of the Cattell Award for early career contributions to multivariate psychology from the Society of Multivariate Experimental Psychology.

Ping Yin (pyin@humrro.org) is Senior Staff Scientist at Human Resources Research Organization (HumRRO). She has over 15 years of professional experience in educational research and measurement, especially for large-scale high-stake testing programs. Her area of interest includes computer adaptive testing and various applications of educational measurement and statistics. Her work has appeared in *Educational and Psychological Measurement*, *Applied Psychological Measurement*, and *Journal of Educational Measurement*.

Index

- a priori* analysis: ANOVA 23–24; cluster analysis 48–50; correlation/association measures 60–61; factor analysis 101, 112, 116–119; item response theory 151–152; latent growth curves 179–182; latent variables 206–211; logistic regression 221–223; power analysis 383–384; structural equation modeling 445, 450–451
absolute indices 115, 453
adequacy coefficients 37–39
agreement, interrater reliability and 132–144, 133
Akaike information criterion (AIC) 20–21, 171, 198, 454
 α error rate levels *see* Type I error (α) rates
alpha scores 119–120
analysis of covariance (ANCOVA) 1, 5, 9, 12, 24
analysis of variance (ANOVA) 1–28, 2, 16, 178–186, 349; correlation/association measures 55–57, 68; G theory 123, 128, 131; interrater reliability 136–142; log-linear analysis 235, 238–239; MC simulations 274–275; moderation 253; multitrait–multimethod matrices 333; multivariate 15–25, 348–361; one-way design 1, 5, 11, 136–142, 312, 368–369
ANCOVA *see* analysis of covariance
appropriateness aspects 30–31, 147, 160–161
approximation methods: ANOVA 15, 20–22, 25; nonparametric statistics 373–374; root mean square error 115, 343, 450–451, 453
articulation (language) 2, 13–14, 191–192, 404–409, 456
artifact correction procedures 266–267
association measures 55–71, 56, 231–232, 356, 366, 370–373, 378
assumption aspects 18–21, 95, 166, 169, 188, 204, 210, 339–340, 451–452, 461; ANOVA 2, 9–10; correlation/association measures 64–67; mediation 251–252; moderation 254; multilevel modeling 304; multiple regression 318–319, 325–329; nonparametric statistics 371–373; propensity scores 391
attrition 9, 27, 187, 412, 419
average treatment effects (ATE) 390–392

B effect sizes 25–26
beta β coefficients 96
baselines 419–431
basis curves 178, 185
Bayesian estimation 127, 151, 341–343
Bayesian information criterion (BIC) 20–21, 171, 198, 454
between-groups analysis of variance 1–14
BIC *see* Bayesian information criterion

binary covariates 390, 394
binary logistic regression models 214, 219–220
bivariate correlations 31, 33
blocking variables 1–3, 6–7
blockmodels, social networks 440
Bonferroni procedures 12, 23–24, 350
bootstrapping 20, 25–26, 479–480
boundary values 174, 188–189
buffers, moderation analysis 255–256

calibration 159–160, 366
canonical correlation analysis (CCA) 29–41, 30
canonical correlation coefficients 32–39
canonical functions/variates 29–40
categorical data analysis 164–175, 362, 365–373, 376–378
categorical predictors 223
categorical variables 1–5, 9–13, 164–165, 235–237
causal inferences: correlation 69; latent growth curves 191–192; latent transition analysis 195; measures of association 69; multilevel modeling 310; propensity scores 388–395; research design 402–415; structural equation modeling 445–448, 456
CCA *see* canonical correlation analysis
censoring 86–94, 97
centering 300–301, 323, 357–358
chain-referrals 471
chi-square statistic 157, 160, 169–171, 241–242, 454, 463
classical test theory 123–125, 130, 137–138, 145
classification procedures 42–54, 147–149, 164–177, 232–240
class parameters 205–206
class proportions 168–169, 172–175
cluster analysis 42–54, 43, 164–167, 202–207, 210–211, 292–310, 440
CMLE *see* conditional maximum likelihood estimation
CO *see* cumulative odds
coding studies 132–135, 223, 262–263
coefficients: canonical correlation 32–39; correlation 32–39, 55–56, 65–66, 135–140, 305–306, 369–370; event history/survival analysis 96; factor analysis 102–108, 110–112, 117–120; G theory 130; interrater reliability 132, 135–142; item response theory 161–162; logistic regression 225–226; multiple regression 313–329; single-subject research 426–428; structure 35–37, 102–108
cohort sequential design 27
collinearity 229–230

- common factor analysis *see* factor analysis
 communality 38, 102–103, 106–110
 comparison methods: ANOVA 16, 23–25; factor analysis 105; latent growth curves 190; latent transition analysis 200; latent variables 211; log-linear analysis 241–243; MANOVA 360; multitrait–multimethod matrices 336–338; nonparametric statistics 376; *post hoc* comparisons 12, 23–25, 83–84, 211, 243–245, 376; structural equation modeling 454, 463–464
 competing risks 88, 91–92
 complementary log-log transformation 219–220
 component analysis 98, 104, 417, 426–427
 composite score quality 110–111, 119–120
 composite variables 348–360
 comprehension knowledge 286
 computer programs *see* software packages
 conditional independence 166
 conditional maximum likelihood estimation (CMLE) 150–152
 conditional probabilities 168–169, 172–174, 195–197
 conditional process analysis 248, 257–259
 conditional tests 353–354
 condition selection, MC simulations 271–272
 confidence intervals 72–85, 73; ANOVA 2, 9, 13, 17, 25–26; multiple regression 315–318, 321, 324–327; nonparametric statistics 371, 377; structural equation modeling 450–451, 453
 confirmatory analysis: factor analysis 98–101, 111–120, 331–339, 343–345, 445–448, 454; latent class analysis 165–168; latent variables 204; log-linear analysis 235–236, 247; multidimensional scaling 277–282, 290; multiple regression 315–316, 320, 329
 confounders 96, 391, 395
 consistency reliability 399, 410
 constructs: definitions 133–134; factor analysis 98–120; interrater reliability 133–143; item response theory 145–149; MANOVA 351, 357–360; validity 400–402, 414–415
 contingency tables 239–240
 continuation ratios 220, 229
 continuous variables 1–5, 9–13, 390
 contrasts 12, 16–19, 23–24, 244–245
 control variables 186, 450
 convenience survey sampling 471
 convergence 188–189, 331, 334, 339–346, 452
 correlation: association measures 55–71, 56; canonical correlation 29–41; coefficients 32–39, 55–56, 65–66, 135–140, 305–306, 369–370; factor analysis 101–119; intraclass correlation 135–140, 305–306; trait–correlation 335–344; zero-order 35, 66, 102, 107
 covariance: ANOVA 1, 5, 9, 12, 15–21, 24–26; correlation/association measures 55, 65, 67; factor analysis 98–120; latent growth curves 178–191; latent variables 206–210; multilevel modeling 304–305; structural equation modeling 457–466
 covariates: event history/survival analysis 90–96; latent class analysis 165–167; propensity scores 388–395
 coverage errors 467–471, 481
 Cox regression 88–97
 criterion-related validity 399–400
 Cronbach’s coefficient 110–111, 119
 cross-classification 239–240, 292, 303
 crossed/crossing variables 1, 4–8, 27
 cross-level data/interactions 62, 293–297, 306, 310
 cross-sectional data 250, 277–290
d effect size 25–26, 75–81
 data: accessibility 94, 189, 211, 394, 452; analysis 149–150, 214–216, 219–221, 226–232, 250–251, 430–431; cleaning 67–68; collection 237–238, 426–427, 437–438, 467–481; generation 272–274; normality 62–63; presentation 39–40, 164–175, 189, 428–430, 452, 461–462; processing 438; reduction 42, 44; screening 103, 113–114; transformations 19–20, 67–68
 DDA *see* descriptive discriminant analysis
 decay studies 277, 279, 288, 290
 decision studies 126–131, 141–142
 deflation 476–477
 degrees of freedom 10–11, 15, 20–22, 25
 dendograms 50
 density, social networks 435
 dependent effect sizes, meta-analysis 264–265
 dependent variables (DV): ANOVA 1–5, 9–18, 24; confidence intervals 77; effect sizes 77; MC simulations 272, 274–275; mediation 248–256; partial correlations 67; single-subject research 425–427
 descriptive discriminant analysis (DDA) 348–353, 358
 descriptive statistics 51, 62, 267, 305, 342, 366, 434–442
 design *see* research design
 design effect 61, 306, 476–478
 DIF *see* differential item functioning
 difference-based hypothesis 4–5
 difference variables 1–6, 10–13
 differential item functioning (DIF) 159, 197
 differential weighting 45–46, 93, 339
 dimensionality analysis 155, 280–281
 discrete-time methods 88–89
 discriminant analysis 348–353, 358–359
 discriminant validation 331
 discrimination expectations 208–209
 distance measures 277–291
 distance metrics 281–284
 distribution-free methods *see* nonparametric statistics
 Dunn–Bonferroni procedure 12, 23–24
 duration analysis *see* event history analysis
 DV *see* dependent variables
 EAP *see* expected a posteriori
 editing errors 467, 481
 effect sizes 72–85, 73; ANOVA 2, 9, 13, 17, 25–26; canonical correlation 34–35; correlation/association measures 68–69; log-linear analysis

- 245–246; mediation 252; meta-analysis 72, 75–78, 83–84, 260–268; multilevel modeling 310; multiple regression 315–318, 321; nonparametric statistics 366; power analysis/statistical power 381–387; single-subject research 429–430
- elimination methods 106
- EMS *see* expected mean square
- entropy 171
- EPI *see* Eysenck Personality Inventory
- equivalence tests 4–5, 13
- errors: ANOVA 2, 4–12, 19–24, 26, 350–356, 359; confidence intervals 72, 76–82; correlation analysis 55, 59–66; covariance structures 304–305; event history/survival analysis 91–93, 96–97; factor analysis 100–106, 114–120; G theory 123, 127–130; interrater reliability 132–133, 136–144; latent growth curves 185–191; latent transition analysis 196–199; latent variables 204, 210–211; logistic regression 217–219, 221–223, 226–233; mediation 250–252, 258; multilevel modeling 292–310; multiple regression 313, 316–329; power analysis/statistical power 382, 385–387; structural equation modeling 449–459, 462–464; survey sampling 467–473, 476–481
- estimation methods: confidence intervals/effect sizes 76–77; factor analysis 104, 113–115; G theory 123–131; item response theory 145–162; latent growth curves 178, 188–191; logistic regression 220–224; structural equation modeling 445, 448–459, 463–465; event history/survival analysis 86–97, 87
- evidence of validity 400
- exacerbators, moderation analysis 255–256
- exact methods 238, 373–374
- expected mean square (EMS) 128
- expected person response function 160–161
- expected a posteriori (EAP) 151
- expected values, log-linear analysis 240–241, 247
- experimental design 76–77, 402–415, 419–431
- explanatory variables 235–237, 313–323, 327–329, 348–350, 355–356, 364–371
- exploratory analysis: factor analysis 98–108, 111–114, 120, 207; latent class analysis 165–168, 173; latent variables 204–210; log-linear analysis 235–257, 247; multidimensional scaling 277–281; repeated-measures ANOVA 25; social networks 434–438; structural equation modeling 445–446, 452–455
- extensions, logistic regression 214–216, 219–221, 226–232
- external validity 402, 405, 408, 414–415
- extraction methods 104, 109
- extraneous variability 411–113
- extrapolation, propensity scores 388, 394
- F tests 7–11, 15
- facets: G theory 125–126; interrater reliability 133, 138–142
- factor analysis 98–122, 99–100, 207, 331–339, 343–345, 445–448, 454
- factorial design/experiments 1–11, 16–17, 23, 355
- factorial invariance 180
- failure time analysis *see* event history analysis
- FIML *see* full information maximum likelihood
- final model/solution aspects 10–11, 49–53, 282–283, 286–287, 462–463
- Fisher exact tests 368, 372, 378
- Fisher's least significant difference 11–12
- fishing expeditions 55–57
- fit aspects: event history/survival analysis 95–96; factor analysis 115–118; goodness-of-fit 78, 169–171, 198, 367, 371–372, 377, 463–464; item response theory 154–161; latent class analysis 169–171; latent growth curves 178, 182, 185–191; latent transition analysis 197–199; latent variables 205–212; logistic regression 214, 217–221, 224–232; MC simulations 274; mixture models 207; multidimensional scaling 280–281; multilevel modeling 301–303, 307; multitrait–multimethod matrices 341–345; nonparametric statistics 367, 371–372, 377; repeated-measures ANOVA 20–21; social networks 443; structural equation modeling 450–456, 463–465
- fixed effects 1–8, 11, 295–303, 306–307, 310
- follow-up tests 1–2, 5, 11–13, 21–24
- Friedman test 369, 373, 378–379
- fringeliers 64
- full information maximum likelihood (FIML) 113, 301, 322, 340–341, 351; latent growth curves 187–188
- function coefficients, canonical correlation 32–39
- future analysis: cluster analysis 53; interrater reliability 143; multitrait–multimethod matrices 346; power analysis/statistical power 386–387
- fuzzy clustering 48
- G^2 test 21, 34, 155–156, 169–171, 227–228, 241–242, 454
- generalizability coefficients 130, 140–142
- generalizability theory (G theory) 123–131
- generalized estimating equations (GEE) 91
- generalized/general linear models 10, 18–20, 29, 55, 214–220
- generalized least squares (GLS) 114, 265
- generalized partial credit 149
- goodness-of-fit 78, 169–171, 198, 367, 371–372, 377, 463–464
- graphic techniques: canonical correlation 32; cluster analysis 51–52; confidence intervals 79–81; event history/survival analysis 94–95; latent transition analysis 195, 199; moderation analysis 254–255; multiple regression 325–326; repeated-measures ANOVA 21, 25
- group differences, MANOVA 350–359
- grouping objects/group membership 42–53
- growth curve/trajectory modeling 15–17, 21, 178–192, 293, 301–303
- growth studies, multidimensional scaling 277–280, 288–290
- G theory 123–131, 124
- guessing parameters 148, 151–152, 160–161, 166

- H coefficient 119–120
 hazard analysis *see* event history analysis
 hazard functions/ratios 94–96
 heterogeneous populations 202–213
 hierarchical cluster analysis 48–50
 hierarchical data 61–62, 292–312
 hierarchical design, ANOVA 1–2, 7–8
 hierarchical linear models *see* multilevel modeling
 hierarchical regression 319–320
 H-L *see* Hosmer–Lemeshow test
 homogeneity: ANOVA 5, 8–11, 15–17, 20–21, 25–26; meta-analysis 267; nonparametric statistics 370, 373, 378
 honestly significant difference (HSD) 12
 Hosmer–Lemeshow (H-L) test 231–232
 hypothesis: ANOVA 3–7, 10–13, 16–17, 21–26; cluster analysis 49; correlation/association measures 57, 60–61, 66; factor analysis 98–101, 109, 112–119; latent class analysis 165–172; latent growth curves 178–185, 189–191; latent transition analysis 193–198; logistic regression 217, 224, 227–228, 231–233; log-linear analysis 235–247; MC simulations 271; mediation 248–249; multidimensional scaling 277–279, 283, 288–290; multilevel modeling 294, 298; multiple regression 315–316, 320, 327; structural equation modeling 445–448, 453–459, 463–465
- ICC *see* intraclass correlation coefficients
 identification: latent class analysis 169; latent growth curves 185, 188–189; latent transition analysis 196–197; logistic regression 222–223; log-linear analysis 235–236; multitrait–multimethod matrices 338–339; structural equation modeling 452
 imputation 22–23, 93, 153, 187, 301, 351
 independence of observations 61–62, 292, 476–477
 independent variables (IV) 1–5, 9–13, 67, 248–257, 425–427
 indices: of association 55, 66, 68–69; cluster analysis 51–52; dissimilarity 170; factor analysis 115–120; item response theory 155–156; latent growth curves 189; mediation 252; meta-analysis 266; multitrait–multimethod matrices 343; relative risk 66; reliability 398–399; structural equation modeling 453, 463–464
 individual difference variables 1, 3–6
 individual parameter tests 227–228
 inference 55–69, 133–134, 315–319, 325–328, 366–373, 434–444
 instrumentalist questions 404
 instrument descriptions 150, 159–160, 398
 integrating covariates 208
 interaction effects, ANOVA 11–12, 23–24
 intercepts/slopes 178–191, 254–256, 302–303, 457, 463
 internal consistency reliability 399, 410
 internal validity 402, 405–415
 internal variables 44–47, 51
 interobserver agreements (IOA) 426–428
 interpretation: canonical functions/variates 37; confidence intervals 81–83; effect sizes 81; factor analysis 108–111, 117–120; interrater reliability 142–143; latent growth curves 191–192; latent variables 211–212; logistic regression 228; mediation 252, 259; moderation analysis 255–256, 259; multidimensional scaling 290
 interrater reliability and agreement 132–144, 133
 interval censoring 89–90, 94
 intervention effects: power analysis/statistical power 382–383; single-subject research 417–431
 intraclass correlation coefficients (ICC) 135–140, 305–306, 477
 invariance 159, 205–206, 457–465
 IOA *see* interobserver agreements
 IRT *see* item response theory
 Irwin–Fisher test 368, 372, 378
 item generation 134
 item measurement properties 204–206
 item response theory (IRT) 145–163
 item to item reliability 399
 IV *see* independent variables
 jackknife repeated replication 479
 joint maximum likelihood estimation (JMLE) 149–152
 K1 “rule,” 104–105
 k-means 48–50
 Kaplan–Meier estimation 86, 92–95
 kappa coefficient 132, 135, 142–143
 Kendall’s tau 65, 369–370, 373, 378
 key elements *see* desiderata/key elements
 knots 329
 Kruskal–Wallis test 368–369, 373, 376–379
 kurtosis 19, 103, 114, 160–161, 340, 451–452
 label switching 174
 Lagrange multiplier tests 453–454
 language (articulation) 2, 13–14, 191–192, 404–409, 456
 latent class analysis (LCA) 164–177, 165
 latent constructs 98, 101–104, 110–111, 145
 latent factors 448–449, 454
 latent growth curve modeling (LGM) 15–17, 21, 178–192, 179
 latent Markov model 196
 latent space models 440
 latent trajectory modeling 15–17, 21, 178–192, 179
 latent transition analysis (LTA) 193–201, 194
 latent variables: item response theory 145–148, 154–155, 158–159; latent class analysis 165, 168; latent growth curves 178–184, 188–191; mixture models 202–213, 203; multitrait–multimethod matrices 333–335, 339, 341, 344; structural equation modeling 445–465
 Latin squares 8
 LCA *see* latent class analysis
 least significant differences (LSD) 11–12
 least-squares: generalized least squares 114, 265; item response theory 150–151; meta-analysis 265;

- ordinary least squares 250–251, 258, 318–319;
repeated-measures ANOVA 19–21, 25–26
- left censoring 89–90
- leveraging points 230, 325
- likelihood estimation 89, 92–96, 150–152, 209,
238–239; *see also* maximum likelihood estimation
- likelihood ratio statistic test 21, 34, 155–156,
169–171, 227–228, 241–242, 454
- linear discriminant functions 23, 348
- linearity, multitrait–multimethod matrices 333–336
- link functions 218–220
- loadings, factor analysis 102, 105–109
- logarithmic transforms 19–20
- logistic mediation 249
- logistic regression (logit) 86, 89–90, 94–96, 149,
214–234, 215, 250
- log likelihood 155–156
- log-linear analysis 235–247, 236
- log-log models 89–92
- log odds 149
- longitudinal analysis 180, 277–281, 288–290, 293,
297, 301–307
- LSD *see* least significant differences
- LTA *see* latent transition analysis
- m* methods 331, 342
- McNemar's test 369, 373, 378
- main-effects follow-up/tests 11–12, 354–355
- manifest variables 164–170, 332–346
- Mann–Whitney test 368, 372, 378
- MANOVA *see* multivariate analysis of variance
- MAR *see* missing at random
- marginal maximum likelihood estimation (MMLE)
150–152
- marginal tests 242–243
- margin of error (MoE) 72, 76–77, 80, 82
- matching methods 388–395, 411–412
- mathematical latent class methods 168–169
- maximum likelihood estimation: event history/
survival analysis 89–94; factor analysis 104,
113–114; item response theory 150–152; latent
class analysis 169; latent growth curves 186–187;
logistic regression 220–224; MANOVA 351;
multilevel modeling 298, 301–303; multiple
regression 322; multitrait–multimethod matrices
340–341; repeated-measures ANOVA 22;
structural equation modeling 450–451, 461–462
- MC *see* Monte Carlo simulations
- MCAR *see* missing completely at random
- MCP *see* multiple comparison procedures
- MDS *see* multidimensional scaling
- means: ANOVA 2–3, 7–13, 16, 19–26; cluster
analysis 48–50, 52; latent growth curves 178–182,
185–191; structural equation modeling
457–466, 458
- measured variables 33, 101–102, 112–113, 178–186,
445–456
- measurement errors 326–327, 467–472, 481
- measurement procedures: G theory 124–125;
latent variables 204–206; multilevel modeling
299; multitrait–multimethod matrices 333–335;
nonparametric statistics 365; structural equation
modeling 448
- measures of association 55–71, 231–232, 356, 366,
370–373, 378
- mediation 248–253, 249, 257–259, 324–325
- membership aspects 42–53, 164–168, 171, 175
- meta-analysis 72, 75–78, 83–84, 260–268, 261
- metrics: item response theory 161–162; latent
growth curves 178–181; multidimensional scaling
281–284; transformation 161–162
- minimum average partial procedure 104–105
- missing at random (MAR) 22, 113, 153, 321
- missing completely at random (MCAR) 22, 153, 321
- missing data/observations: ANOVA 2, 9–10, 21–24,
27; cluster analysis 46–47; correlation/association
measures 63; event history/survival analysis
93–96; factor analysis 113; item response theory
153–154; latent growth curves 178, 187–189;
latent variables 210; logistic regression 224–225;
meta-analysis 265–266; multidimensional
scaling 282–283, 289–290; multilevel modeling
301; multiple regression 321–322; multitrait–
multimethod matrices 340–341; propensity scores
393; social networks 438–439; structural equation
modeling 451, 461–462
- missing not at random (MNAR) 22, 321
- mixed effects models *see* multilevel modeling
- mixed models 1, 7, 15–26, 91, 202–213, 203
- MMLE *see* marginal maximum likelihood estimation
- MNAR *see* missing not at random
- model alignment, multilevel modeling 294–295
- model-based methods: cluster analysis 42; latent
class analysis 165–166; logistic regression
228–230; log-linear analysis 238–239; social
networks 439–444; survey sampling 478
- model comparisons *see* comparison methods
- model fit *see* fit aspects
- moderation analysis 248–249, 249, 253–259,
322–324
- MoE *see* margin of error
- Monte Carlo (MC) simulations 269–276, 270, 299
- MS SEM *see* multiple sample structural equation
modeling
- MTTM *see* multitrait–multimethod matrix analysis
- multidimensionality, item response theory 158
- multidimensional scaling (MDS) 277–291, 278
- multi-element single-subject research 424
- multi-factor MANOVA designs 353–355
- multilevel modeling 292–312, 293–294, 329
- multinomial data 214–216, 219–221, 226–232, 237
- multiple baseline design 420–425, 429
- multiple comparison procedures (MCP) 11–12, 16,
23–24
- multiple outcome measures 351
- multiple probe design 420–423, 429
- multiple regression 15, 18–21, 24–26, 313–330, 314
- multiple sample structural equation modeling (MS
SEM) 457–465
- multiple schedule designs 424

- multiple social networks 440–442
 multiple testing strategies 23–24
 multiplicity problem 316
 multisample covariance 457–466
 multitrait–multimethod (MTMM) matrix analysis
 331–347, 332
 multivariate analysis 66, 277–291
 multivariate analysis of variance (MANOVA) 15–25,
 348–361, 349
 multivariate variable relations 29–31, 34–35, 39
 naming factors 108–109
 narrative synthesis, meta-analysis 260–261
 nesting 6–9, 21, 26, 61–62, 95–96, 292–294, 307, 450
 network analysis 434–444
NHST *see* Null Hypothesis Statistical Testing
 nodes, social networks 434–442
 non-experimental studies 388–395, 402–408, 413–414
 non-hierarchical methods 48–50
 non-model-based cluster analysis 42, 46
 non-orthogonality 9
 non-parametric . . . : multiple regression 328; single-subject research 431; statistics 362–379, 363
 non-random relationships 1–2, 5, 10–13
 non-response errors/sources 467, 472–473
 normal distribution 3, 9–10, 15, 18–21, 24–26,
 62–63, 318
 normality of data 62–63
 null hypothesis: ANOVA 7, 10–13, 16–17, 21–24;
 correlation/association measures 60, 72–76,
 80–84; latent growth curves 189–190; latent transition analysis 198; multiple regression 315–316, 320, 327; power analysis/statistical power 380–387
 Null Hypothesis Statistical Testing (NHST) 60,
 72–76, 80–84
 object classification, cluster analysis 42–53
 observation periods, event history 88
 observer ratings 132–144
 odds: log-linear analysis 240–241; ratios 55, 66, 218,
 226, 240–241
 OLS *see* ordinary least squares
 omega scores 119–120
 omnibus tests 10–12, 16–17, 317, 321, 354–357, 360,
 376–378
 one-sample Wilcoxon test 368, 372, 377–378
 one-way ANOVA design 1, 5, 11, 136–142, 312, 368–369
 Open Science 72–74, 77, 84
 operationalizing variables of interest 57–58
 option response functions (ORF) 157
 ordinal data 214–216, 219–221, 226–232
 ordinary least squares (OLS) 250–251, 258, 318–319
 ORF *see* option response functions
 origin time 88, 180
 outcome measures: ANOVA 1–6, 9–10, 13–18,
 23, 27; log-linear analysis 247; MANOVA 351;
 nonparametric statistics 377–378; power analysis/statistical power 382–385; propensity scores 388–395
 outliers: cluster analysis 46–47; correlation/association measures 64; factor analysis 103, 114; latent growth curves 178, 187–189; multitrait–multimethod matrices 340–341; structural equation modeling 451, 461–462
 overdispersion 226–227
 overlaps 390, 394, 429–431
P tests 18
p values 68, 91, 96, 317
PAMS *see* Profile Analysis via Multidimensional Scaling
 panel analysis 193
 parallel analysis 104–105
 parallel odds 220, 229
 parallel replication 124–125, 129
 parameter estimates/values: factor analysis 106–107,
 118–119; latent transition analysis 199–200; latent variables 210–211; logistic regression 220–224, 227–228; multilevel modeling 303; multiple regression 320–321; multitrait–multimethod matrices 331, 338–345; structural equation modeling 445, 448–459, 463–465
 parameter invariance 159
 parameter logistic models 148
 parametric event history/survival analysis 89–90
 parametric single-subject research 417, 431
 parametric statistical procedures *see* analysis of variance
 parsimony-adjusted indices 115, 453
 partial correlations 66–67
 partial credit models 149
 partial likelihood estimation 89, 93–95
 partial tests, log-linear analysis 242–243
 participants: ANOVA 2–3, 6–9, 15–18,
 22, 25–27; multidimensional scaling 280;
 research design 409; single-subject research 417–418
 partitioning, cluster analysis 49–50
 path diagrams: canonical correlation 31–32; latent growth curves 181–184, 191; latent variables 206–207; multitrait–multimethod matrices 335–336; structural equation modeling 445–448, 455, 458–459, 464
 pattern coefficients 102, 104–108
 Pearson product-moment correlation coefficient 32–33, 55–56, 65–66, 103, 169–170, 369–370
 perceived effectiveness, research design 404
 percentage of variance 104, 107–108, 119, 384–385
 permutation tests 362
 person location estimates 161
 person response function 160–161
 Pillai–Bartlett trace 25
 π^* , latent class analysis 170
 point-biserial correlation 65–66
 point estimates 72–79, 82
 Poisson sampling 237–238
 population aspects: MANOVA 348, 351–354;
 multiple regression 313–322, 327–328; survey sampling 468–469

- post hoc:* comparisons 12, 23–25, 83–84, 211, 243–245, 376; model modifications 116–118, 343–344; probing 258
- power analysis 17–19, 34, 60–61, 245–246, 380–387, 381
- precision, point estimates 72–77, 82, 478
- prediction errors 307–310, 317–320, 326–327
- predictive ability, multilevel modeling 307–310
- predictors: event history/survival analysis 87, 90, 94–96; logistic regression 214–233; multiple regression 313–29
- pre-experimental research designs 405
- preliminary analysis, MANOVA 352–353
- preregistration research plans 77–78
- presentation aspects: canonical correlation 39–40; latent class analysis 164–175; latent growth curves 189; latent variables 210–211; MC simulations 275; multilevel modeling 296–297, 307; power analysis/statistical power 385–386; reliability/validity 400–401; single-subject research 428–430; structural equation modeling 445, 452, 461–462
- previous research aspects: effect sizes/confidence intervals 74–75; factor analysis 100–101, 111–112; MC simulations 270–271; multidimensional scaling 279; power analysis/statistical power 386–387; research design 403–404
- primary sampling units (PSU) 470–471, 474–480
- probability: between-groups ANOVA 7, 10–13; latent class analysis 168–169, 172–174; latent transition analysis 195–197; propensity scores 388–395; proportionate to size sampling 470–471, 475; of success 214–228, 231–232; survey sampling 470–476
- probe design 420–423, 429
- probit regression model 219
- process analysis models 257–259
- process errors 467, 481
- Proctor model 168–169
- product-moment correlation coefficient 32–33, 55–56, 65–66, 103, 369–370
- Profile Analysis via Multidimensional Scaling (PAMS) 285–286
- propensity scores 388–395, 388–389
- proportional hazards assumption 95
- proportional odds 220, 229
- proportional reduction in prediction error/variance statistics 308–309
- proportional similarity 370
- proportionate to size sampling 470–471, 475
- proportions, latent class analysis 168–169, 172–175
- proximity measures 47–48, 51, 288
- pseudo-experimental research designs 405
- pseudo-guessing/chance parameters 148, 152
- PSU *see* primary sampling units
- publication bias 266
- Q3 statistic 158
- Q-factor analysis 287–288
- Q test 369
- quality aspects 171–172, 264
- quasi-experimental research 402–408, 411–412
- quasi-F* tests 7
- quasi-separation 229–230
- question formulation 74, 419–425, 469
- random effects 178, 209; *see also* multilevel modeling
- random errors 204, 210–211
- random factors, ANOVA 1–22
- randomized block design 1–3, 6–7
- randomized research designs 405, 424
- random measurement errors 326–327
- random replication 124–125, 129
- rank-based methods 19–20, 65, 362, 365–379
- Rasch modeling 147–151, 157–160
- rater-to-rater reliability 399
- ratings, interrater reliability 132–144
- realist research questions 404–405
- reciprocity 435
- redundancy coefficients 37–39
- regression: coefficients 78, 313–329; discontinuity 406–407; event history analysis 86–97; logistic regression 86, 89–90, 94–96, 149, 214–234, 215, 250; multiple regression 15, 18–21, 24–26, 313–330, 314; regressor criterion/justification 315–316
- rejected/retained models 198–199
- relative error variance 129–130
- relative risk 66
- reliability 397–401, 397; correlation analysis 58–59, 65–67; G theory 124–125, 129–131; interrater reliability 132–144; research design 409–413; structural equation modeling 454; *see also* event history analysis
- repeated events 88, 91, 93–94
- repeated-measures analysis of variance 15–28, 178–186
- replicability: cluster analysis 42, 46, 50–53; confidence intervals 83–84; effect sizes 83–84; G theory 124–125, 129; interrater reliability 132, 137–140, 143–144; repeated replication 479–480; single-subject research 425–426, 431; survey sampling 479–480
- research design 402–416, 403; ANOVA 5–6, 15, 25; canonical correlation 39; correlation analysis 57; factor analysis 100–101, 111–112; G theory 125–126; MANOVA 349–351, 360; MC simulations 269–271; multiple regression goals 314–315; preregistration plans 77–78; questions 57, 269–271, 349–351, 360, 419–425, 469; single-subject research 417–431; survey sampling 469–471, 476–481
- residual diagnostics 230–231
- residual multilevel modeling 304
- residual variances 188–189
- response data 149–150
- response interruption and redirection (RIRD) 420
- response probability 214–228, 231–232
- response rates 472–473
- response theory 145–175
- response variables 216–217, 364–371

- restricted maximum likelihood (REML) 303, 307, 322
- result presentation 275, 385–386, 400–401, 445
- retained latent transition models 198–199
- reversal design 419–420, 424–425, 429
- right censoring 86–89
- RIRD *see* response interruption and redirection
- robust procedures, ANOVA 2, 9–11, 19–20, 25
- root mean square error of approximation (RMSEA) 115, 343, 450–451, 453
- rotational methods 105–108, 281
- sampling/sample size: cluster analysis 46, 50–53; correlation/association measures 59–64, 68; errors 467, 470, 478, 481; event history/survival analysis 92–93; factor analysis 101–122; frames 469–470; latent class analysis 167–168; latent growth curves 186–187; logistic regression 220–221; log-linear analysis 237–238; meta-analysis 262–263; multidimensional scaling 280–281; multilevel modeling 297–309; multiple regression 320–321; multitrait–multimethod matrices 338, 343–346; power analysis/statistical power 380–387; propensity scores 390–391, 394; repeated-measures ANOVA 17–22, 26; single-subject research 427; structural equation modeling 450–454, 457–465; survey sampling 467–481
- saturated log-linear models 242
- scales: factor analysis 110–111; interrater reliability 132–135, 140–143; latent growth curves 180; multidimensional scaling 277–291
- scores: factor analysis 109–111, 114, 117–120; interrater reliability 132–143; propensity scores 388–395
- screening data 103, 113–114
- SD *see* standard deviation
- SE *see* standard errors
- selection techniques: interrater reliability 135; research design 404–409; social networks 441; survey sampling probabilities 473–476
- SEM *see* structural equation modeling
- semi-parametric procedures 89–90
- semipartial correlations 66–67
- sensitivity, power analysis/statistical power 381–382
- separation, logistic regression 229–230
- sequence effects 419, 424
- SES *see* socio-economic status
- setting aspects, single-subject research 417–418
- significance levels/tests: ANOVA 1–2, 9–13, 17, 20, 23–25; canonical correlation 34; factor analysis 118; latent class analysis 169–170; multitrait–multimethod matrices 331; power analysis/statistical power 383–387; structural equation modeling 464; *see also* statistical significance
- sign tests 367–368, 371, 377
- simple effects, ANOVA 12
- simple random sampling (SRS) 470–471, 476
- Simpson's paradox 242–243
- simulation studies 49, 269–276, 298–299
- single-factor designs 353–355
- single-subject research 417–433, 418
- skewed distribution 19–20, 67–68, 103, 212, 460–461
- skipped correlation coefficients 65–66
- slopes/intercepts 178–191, 254–256, 302–303, 457, 463
- Sobel's test 250–252
- social network analysis 434–444, 436
- socio-economic status (SES) 293–297, 306
- sociograms 434–435, 442
- software packages: ANOVA 9, 12, 17–19, 24–26; cluster analysis 42, 51; event history/survival analysis 89–96; factor analysis 99–100, 103–104, 107–108, 113–115; G theory 123, 127; item response theory 145–146, 151–157, 160–161; latent class analysis 169, 171, 174; latent growth curves 178, 188–191; latent transition analysis 197–199; latent variables 209–210; logistic regression 222; log-linear analysis 235–237, 243; MANOVA 353; multilevel modeling 303; nonparametric statistics 366; power analysis/statistical power 385–386; propensity scores 393; structural equation modeling 445, 449–452, 455–457, 461; survey sampling 474–481
- sparseness 245, 435
- Spearman's correlation 65–66, 369–370, 373, 378
- specification procedures: G theory 126; item response theory 147–149; latent growth curves 190–191; multilevel modeling 296–297; propensity scores 393–394; structural equation modeling 445, 452–459, 462–464
- spherical covariance structures 20, 25
- spline regression models 328–329
- split-plot design 17
- SRS *see* simple random sampling
- standard deviation (SD) 75–80, 316–318, 323, 327, 382–383
- standard errors (SE): confidence intervals 79–81; event history/survival analysis 91–93, 96–97; factor analysis 102–104, 114–120; latent transition analysis 196–199; latent variables 204, 210–211; logistic regression 217–223, 226–233; mediation 250–252, 258; multilevel modeling 292–299, 306–307; multiple regression 322, 326–327; power analysis/statistical power 382; survey sampling 476–481
- standardized regression coefficients 225–226, 318, 326
- state models 284–286
- statistical assumptions *see* assumption aspects
- statistical conclusion validity 402, 412–413
- statistical multilevel model presentation 296–297
- statistical power 17–19, 34, 60–61, 245–246, 380–387, 381
- statistical significance 8–9, 24–25, 34, 268, 383–387, 464; *see also* significance levels/tests
- statistical testing 29–30, 33–35, 170–171, 245–246, 380–387
- step wedge design 27
- stratified sampling 470, 478
- strength of association 356

- structural equation modeling (SEM) 445–466, 446, 458; factor analysis 98–99, 113–114, 445–448, 454; latent growth curves 178–191; latent transition analysis 195, 199; multitrait–multimethod matrices 333, 338–341; repeated-measures ANOVA 15, 21, 27
- structure coefficients 35–37, 102–108
- structured means-structural equation modeling 457–465
- study designs *see* research design
- subclassification, propensity scores 392–394
- subscale scores 110–111, 119–120
- summary statistics: canonical correlation 33; event history/survival analysis 94; G theory 127; multitrait–multimethod matrices 342; repeated-measures ANOVA 21, 25; single-subject research 428–430; social networks 434–436, 439–443
- Sums of Squares 9, 375
- survey sampling 434, 437–440, 444, 467–481, 468
- survival analysis *see* event history/survival analysis
- survivor functions 94–96
- systematic replication 425
- systematic survey sampling 470
- t*-tests 350, 380, 387
- tabular presentation 198–199, 307, 462
- targeted effect sizes 317–318, 321
- target metrics 161–162
- taxonomy, item response theory 147–149
- test procedures: ANOVA 1–27; canonical correlation 29–30, 33–35; logistic regression 227–228; log-linear analysis 241–246; MANOVA 353–360; mediation 251; nonparametric statistics 367–373, 376–379; power analysis/statistical power 380–387; single-subject research 419, 430–431; statistical testing 29–30, 33–35, 170–171, 245–246, 380–387; structural equation modeling 454, 463; test characteristic curve equating 162; test-retest reliability 398; *see also* significance levels/tests
- test statistics: ANOVA 10–12, 20, 24–26; event history/survival analysis 92–93; latent transition analysis 198; latent variable mixture models 209; logistic regression 227, 231; log-linear analysis 241–245; MANOVA 353–354; meta-analysis 267; multitrait–multimethod analysis 343; nonparametric statistics 375–379; power analysis 380–381; structural equation modeling 455; survey sampling 478
- theoretical aspects: correlation/association measures 55–57; factor analysis 100–101, 111–112; latent class analysis 165–166; latent growth curves 179–180; latent transition analysis 194–195; logistic regression 217; log-linear analysis 246–247; meta-analysis 260–261; multidimensional scaling 279; multilevel modeling 294–295; multitrait–multimethod matrices 333; social networks 435–437, 443; structural equation modeling 445–448, 458–459, 465
- three-way factorial design 1
- ties 89, 374–376, 434–444
- time-dependent/varying covariates 90–91, 94–95, 186
- time metrics 178, 180–181
- time points 288–289
- time sampling 427
- TLI *see* Tucker–Lewis index
- total characteristic function equating 162
- total scale scores 110–111, 119–120
- training, interrater reliability 135
- traits, multitrait–multimethod matrices 331–347
- trajectory modeling 15–17, 21, 178–192, 293, 301–303
- transformations: cluster analysis 45–46; item response theory 161–162
- transforms, ANOVA 19–20
- transition analysis *see* event history analysis; latent transition analysis
- transparency 73, 325–326
- treatment effects, power analysis 382–383
- trimmed means 10–11, 19–21, 25–26
- Tucker–Lewis index (TLI) 115, 343
- two-phase structural equation modeling 452–453
- two-sample Wilcoxon test 368, 372
- Type I error (α) rates: ANOVA 9–12, 19–23, 26; correlation/association measures 60–61, 66; factor analysis 100; MANOVA 350–355; multilevel modeling 292, 298, 306; multiple regression 316, 320–324, 329; power analysis/statistical power 382, 385–387
- Type II errors 60, 66, 382, 385–387
- unconditional probabilities 195–196
- unitization problem, interrater reliability 134
- univariate analysis 18–19, 21, 66
- universes, G theory 125–130
- unmeasured confounding 391
- unobserved confounders 395
- unobserved factors, structural equation modeling 446–448
- unstandardized coefficients 118–119, 318, 326
- unweighted least squares 150–151
- unweighted pooled covariance 206–207
- utility aspects, cluster analysis 52–53
- V statistic 371
- validation/validity 397–401, 397; cluster analysis 42–46, 50–53; correlation analysis 58–59, 64; latent variables 210–211; multitrait–multimethod matrices 331, 344–346; nonparametric statistics 371–373; repeated-measures ANOVA 16–21, 26–27; research design 402–415; structural equation modeling 454
- variability, single-subject research 429
- variable relationships: canonical correlation 29–40; cluster analysis 42–53; correlation/association measures 55–69; factor analysis 98–120; item response theory 145–159; latent class analysis 164–175; latent growth curves 178–192; log-linear analysis 235–247; mediation 248–253, 257–259; moderation 253–259; multidimensional scaling

- 281–283; multilevel modeling 292–301; multiple regression 316, 319–320; structural equation modeling 445–465
- variance: components 13, 125–131, 141, 295, 298–309; factor analysis 98, 104, 107–108, 119–120; G theory 123–131; inflation 476; interrater reliability 141; latent growth curves 178–191; latent variables 205–206; multilevel modeling 295, 298–309; survey sampling 476–481
- Venn diagrams 126
- visual aspects 286, 325–326, 434–435, 440–442
- Wald z statistic 25–26, 227–228, 454
- weighting/weights: canonical correlation 32–39; cluster analysis 45–48; latent variables 202, 206–207; logistic regression 222; MANOVA 358–359; meta-analysis 264; propensity scores 392–394; survey sampling 473–476
- Welch test 10–11
- Widaman taxonomy 343
- Wilcoxon test 368, 372, 377–378
- Wilks's lambda 25
- Winsorized correlations 65
- Winsorized variances 10–11, 19–21, 25–26
- with-in subjects analysis of variance 15–28
- z -scores 45, 78
- z statistic 25–26, 227–228, 454
- zero-cell problems 229–230
- zero-order correlations 35, 66, 102, 107



Taylor & Francis Group
an informa business



Taylor & Francis eBooks

www.taylorfrancis.com

A single destination for eBooks from Taylor & Francis with increased functionality and an improved user experience to meet the needs of our customers.

90,000+ eBooks of award-winning academic content in Humanities, Social Science, Science, Technology, Engineering, and Medical written by a global network of editors and authors.

TAYLOR & FRANCIS EBOOKS OFFERS:



A streamlined experience for our library customers



A single point of discovery for all of our eBook content



Improved search and discovery of content at both book and chapter level

REQUEST A FREE TRIAL

support@taylorfrancis.com

 Routledge
Taylor & Francis Group

 CRC Press
Taylor & Francis Group