Exploratory factor analysis with SAS

Book · January 2016

1 author:

Jason W Osborne
Clemson University

104 PUBLICATIONS **6,335** CITATIONS

Some of the authors of this publication are also working on these related projects:

Regression and linear modeling book and ancillary materials View project

Best Practices in Exploratory Factor Analysis

Jason W. Osborne, Ph.D., PStat®

Printed and Kindle versions of this book are available through Amazon :

http://www.amazon.com/-/e/B00FCLJQES

Cite: Osborne, J. W. (2014). Best Practices in Exploratory Factor Analysis. Scotts Valley, CA: CreateSpace Independent Publishing. ISBN-13: 978-1500594343, ISBN-10:1500594342.

DEDICATION

This book, as all my work, is dedicated to my family. They are my *sine qua non*.

TABLE OF CONTENTS

The engineering data are used with permission of Dr. Marie Paretti. The data were from a larger study supported by the National Science Foundation under Grant No. HRD# 0936704. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

The Marsh SDQ data were drawn from the National Center for Educational Statistics study: National Education Longitudinal Study of 1988. These data are in the public domain. More information can be found at: http://nces.ed.gov/surveys/nels88/ .

The Geriatric Depression Survey data were drawn from the publicly available data from the Long Beach study. The citation for these data, residing at the ICPSR is:

Zelinski, Elizabeth, and Robert Kennison. Long Beach Longitudinal Study. ICPSR26561-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2011-06-17.

More information on this study is available at:
http://www.icpsr.umich.edu/icpsrweb/NACDA/studies/26561/version/2

The opinions about these data, and the results from these analyses are solely my opinion opinions and interpretations, used for the pedagogical purpose of discussing best practices and techniques around exploratory factor analysis. They should be viewed as having no bearing on the authors of those original studies, the findings from those studies, or just about anything else. Furthermore, where weighting would be appropriate (such as with the Marsh data), I did not apply any weights or compensate for design effects, thus rendering the results not useful for anything other than these examples.

ACKNOWLEDGMENTS

This work is based on several articles originally published in *Practical Assessment, Research, and Evaluation (PARE)*, which can be accessed at http://pareonline.net/. This volume represents work that happened over the course of a decade, much in collaboration with two students from my time at North Carolina State University: Blandy Costello and David Fitzpatrick. Their collaborations have been invaluable impetus to exploring these issues.

I would also like to acknowledge Dr. Lawrence Rudner, and Dr. William Schafer, editors of PARE. These two scholars have been a source of valuable feedback and insight since 2000 when they welcomed me into their group of reviewers. Not long after, they accepted a brief paper on Hierarchical Linear Modeling, gently guiding my work to be much better than it otherwise would have been. When I had the wild idea to begin working on a book several years later, Dr. Rudner encouraged me to dream big. His encouragement has inspired me to do things I might not have attempted otherwise.[1]

Works that were drawn upon for this volume:

Osborne, J. W., & Costello, A. B. (2004). Sample size and subject to item ratio in principal components analysis. *Practical Assessment, Research, and Evaluation, 9*. Online at http://pareonline.net/getvn.asp?v=9&n=11 .

Costello, A. B. & Osborne, J. W. (2005). Exploratory Factor Analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10(7),* 1-9. Online at http://pareonline.net/pdf/v10n7.pdf.

Osborne, J. W., & Fitzpatrick, D. (2012). Replication Analysis in Exploratory Factor Analysis: what it is and why it makes your analysis better. *Practical Assessment, Research, and Evaluation, 17(15),* 1-8. http://pareonline.net/getvn.asp?v=17&n=15.

Osborne, J. W. (in press, expected 2014). What is rotating in exploratory factor analysis? *Practical Assessment, Research, and Evaluation.*

[1] In other words, for those of you who wish I had *not* been encouraged to publish my ramblings about statistical practice, he is primarily to blame. Direct all criticism and hate mail to him…

1 INTRODUCTION TO EXPLORATORY FACTOR ANALYSIS

Exploratory factor analysis (EFA) is a statistical tool used for many purposes. It was originally developed in the early 1900s during the attempt to determine whether intelligence is a unitary or multidimensional construct (Spearman, 1904). It has since served as a general-purpose dimension reduction tool with many applications. In the modern social sciences it is often used to explore the psychometric properties of an instrument or scale. Exploratory factor analysis examines all the pairwise relationships between individual variables (e.g., items on a scale) and seeks to extract latent factors from the measured variables. During the 110 years since Spearman's seminal work in this area, few statistical techniques have been so widely used (or, unfortunately, mis-used). The goal of this book is to explore best practices in applying EFA. We will occasionally survey some poor practices as a learning tool. Let us start first with a brief discussion about the similarities and differences between Principal Components Analysis (PCA) and Exploratory Factor Analysis.

The difference between Principal Components Analysis and Exploratory Factor Analysis.

Much has been written about the differences between these two techniques, and many misconceptions exist about them. One of the biggest misconceptions is that PCA is part of EFA, although they both seem to do the same thing. This misconception probably has modern day roots from at least two factors:

1. Many statistical software packages have PCA as the default extraction technique when performing exploratory factor analysis,
2. Many modern researchers use PCA and EFA interchangeably, or use PCA when performing an analysis that is more appropriate for EFA

Principal components analysis is a computationally simplified version of a general class of dimension reduction analyses. EFA was developed before PCA, thanks to Spearman. EFA was developed prior to the computer age when all statistical

calculations were done by hand, often using matrix algebra. As such, these were significant undertakings requiring a great deal of effort. Because of the substantial effort required to perform EFA with hand calculations, significant scholarship and effort went into developing PCA as a legitimate alternative that was less computationally intense but that also provided similar results (Gorsuch, 1990). Computers became available to researchers at universities and industrial research labs later in the 20th century, but remained relatively slow and with limited memory until very late in the 20th century (about the time I was in graduate school using mainframes at the University[2]). My commentary on PCA is not to sleight these scholars nor to minimize their substantial contributions, but rather to attempt to put PCA and EFA into context for the modern statistician and quantitative researcher. I will therefore focus on EFA, despite the popularity of CFA.

Without getting into the technical details, which are available in other scholarly references on the topic, PCA computes the analysis without regard to the underlying latent structure of the variables, using all the variance in the manifest variables. What this means is that there is a fundamental assumption made when choosing PCA: that the measured variables are themselves of interest, rather than some hypothetical latent construct (as in EFA). This makes PCA similar to multiple regression in some ways, in that it seeks to create optimized weighted linear combinations of variables.

As you can see in Figure 1.1., Principal Components Analysis combines manifest (observed) variables into weighted linear combinations[3] that end up as components.

Exploratory factor analysis, on the other hand, is a group of extraction and rotation techniques that are all designed to model unobserved or latent constructs. It is referred to as common factor analysis or exploratory factor analysis.

EFA assumes and asserts that there are latent variables that give rise to the manifest (observed) variables, and the calculations and results are interpreted very differently in light of this assumption.

You can see this very different conceptual vision in Figure 1.2, below. Factor analysis also recognizes that model variance contains both shared and unique variance across variables. EFA examines only the shared variance from the model each time a factor is created, while allowing the unique variance and error variance to remain in the model. When the factors are uncorrelated and communalities are moderate, PCA can produce inflated values of variance accounted for by the components (Gorsuch, 1997; McArdle, 1990). Since factor analysis only analyzes shared variance, factor analysis should yield the same general solution (all other things being equal) while also avoiding the illegitimate inflation of variance estimates.

Figure 1.1: Conceptual overview of Principal Components Analysis
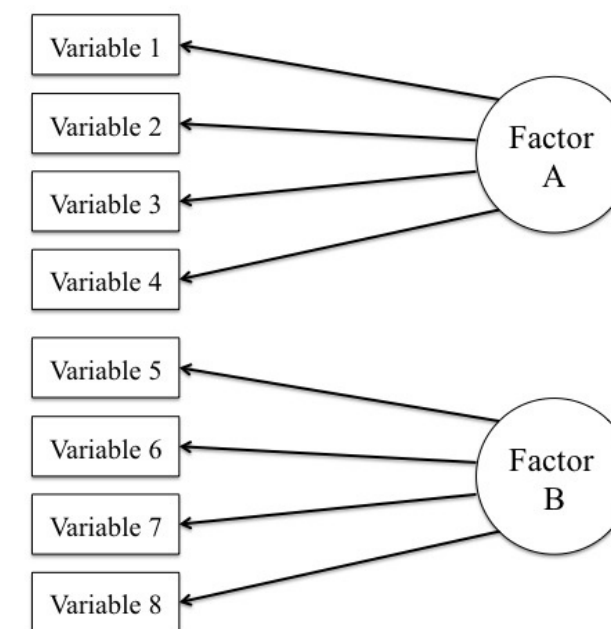
Figure 1.2: Conceptual overview of Exploratory Factor Analysis

---

[2] and walking across campus uphill both ways in the snow to get print outs of my analyses. It was Buffalo before climate change thawed that particular tundra.

[3] Weighted linear combinations means that each variable has a different weight- the factor/component loading that determines how much or little each variable contributes to the composite. This is similar to multiple regression where a variable score is composed of different amounts (regression weights) of each variable in the equation.

There are two other issues with PCA that I will briefly note. First, PCA assumes that all variables are measured without error (an untenable assumption in almost any discipline), whereas EFA has the option of acknowledging less than perfect reliability. Second, PCA parameters are selected in an attempt to reproduce sample, rather than population characteristics (Thompson, 2004).

Thus, we have many similarities between PCA and some important conceptual and mathematical differences. Most authors agree that there is little compelling reason to choose PCA over other extraction methods, and that PCA can be limited and provide biased parameter estimates (including, but not limited to: Bentler & Kano, 1990; Floyd & Widaman, 1995; J.K. Ford, R.C. MacCallum, & M. Tait, 1986; Gorsuch, 1990; Loehlin, 1990; MacCallum & Tucker, 1991; Mulaik, 1990; Widaman, 1993). If one is to seek best practices, one is hard pressed to conclude PCA is ever a best practice. Widman (1993) puts it very bluntly: "principal components analysis should not be used if a researcher wishes to obtain parameters reflecting latent constructs or factors." (p. 263). Unfortunately, it is still the default dimension reduction procedure in many statistical analysis software packages despite it usually not being (in my opinion) the conceptually desirable choice, and having no clear advantage in modern quantitative methodology that I can detect.

This is a topic that arouses passions amongst statisticians, and I have published few papers or given few talks on this topic without someone getting upset at me for taking this position so clearly and unapologetically. So let us sidestep this issue for the moment and summarize: PCA is not considered a factor analytic technique and there is disagreement among statisticians about when it should be used, if at all. More often than not, researchers use PCA when EFA would be appropriate and preferable (for example, see Ford, MacCallum, & Tait, 1986; Gorsuch, 1983; Widaman, 1993).

Steps to follow when conducting EFA

Exploratory factor analysis is meant to be exploratory in nature, and thus it is not desirable to prescribe a rigid formula or process for executing an EFA. The steps below are meant to be a loose guide, understanding that a factor analysis often requires returning to previous steps and trying other approaches to ensure the best outcome. The general pattern of performing an EFA falls into six general steps that will guide the discussion through the rest of the book:

1. Data cleaning
2. Deciding on extraction method to use
3. Deciding how many factors to retain
4. Deciding on a method of rotation (if desired)
5. Interpretation of results
   (return to #3 if solution is not ideal)
6. Replication or evaluation of robustness
   (return to beginning if solution is not replicable or robust)

Step 1: Data cleaning. I have been writing about the importance of data cleaning for about as long as I have been writing about best practices in quantitative methods because without clean data, what follows in almost any analysis is moot. This is another point where passions run high amongst researchers and statisticians. I have a clear position on the issue, having written an entire book on the topic, demonstrating repeatedly how clean data produces results that are better estimates of population parameters, and therefore, more accurate and replicable (Osborne, 2013). Instead of debating the point here, allow me to assert that data filled with errors and /or which fails to meet assumptions of the analysis being performed is likely to lead to poorer outcomes than data that is free of egregious errors and that meets assumptions. We will discuss some other data quality issues later in the book, including the importance of dealing appropriately with missing data.

Step 2: Deciding on an extraction method. An extraction technique is one of a group of methods that examines the correlation/covariation between all the variables and seeks to "extract" the latent variables from the measured/manifest variables.

There are several factor analysis extraction methods to choose from. SPSS has six (in addition to PCA; SAS and other packages have similar options): unweighted least squares, generalized least squares, maximum likelihood, principal axis factoring, alpha factoring, and image factoring. Information on the relative strengths and weaknesses of these techniques is not easily had. To complicate matters further, naming conventions for some extraction techniques are not consistent, leaving it difficult to figure out which method a textbook or journal article author is describing, and whether or not it is actually available in the software package the researcher is using. This probably explains the popularity of principal components analysis – not only is it the default in many statistical packages, but it is one of the more consistent names researchers will see in statistical packages.

A recent article by Fabrigar, Wegener, MacCallum and Strahan (1999) argued that if data are relatively normally distributed, maximum likelihood is the best choice because "it allows for the computation of a wide range of indexes of the goodness of fit of the model [and] permits statistical significance testing of factor loadings and correlations among factors and the computation of confidence intervals." (p. 277). If the assumption of multivariate normality is "severely violated" they recommend one of the principal factor methods; in SPSS this is principal axis factors (Fabrigar et al., 1999). Other authors have argued that in specialized cases, or for particular applications, other extraction techniques (e.g., alpha extraction) are most appropriate, but the evidence of advantage is slim. In general, ML or PAF will give you the best results, depending on whether your data are generally normally-distributed or significantly non-normal, respectively. In Chapter 2, we will compare outcomes between ML and PAF, along with some of the other more common extraction techniques.

Step 3: Deciding how many factors to retain for analysis. This too is an issue that suffers from anachronistic ideas and software defaults that are not always ideal (or even defensible). In this step, you (or the software) decide how many factors you are going to keep for analysis. The statistical software will always initially extract as many factors as there are variables (i.e., if you have 10 items in a scale, your software will extract 10 factors) in order to account for 100% of the variance. However, most of them will be meaningless. Remembering that the goal of EFA is to *explore* your data and *reduce* the number of variables being dealt with, there are several different ways of

approaching the decision of how many factors to extract and keep for further analysis. Our guide will always focus on the fact that extracted factors should make conceptual and theoretical sense, and be empirically defensible. We will explore guidelines for this later in Chapter 2.

Step 4: Deciding on a rotation method and rotating the factors. Rotation is a source of some confusion, leading me to write a paper recently with the goal of describing what exactly rotation is, and what is happening when data are rotated. In brief, the goal is to clarify the factor structure and make the results of your EFA most interpretable. There are several different rotation methodologies, falling into two general groups: orthogonal rotations and oblique rotations. Orthogonal rotations keep axes at a 90º angle, forcing the factors to be uncorrelated. Oblique rotations allow angles that are not 90º, thus allowing factors to be correlated if that is optimal for the solution. I argue that in most disciplines constructs tend to be at least marginally correlated with each other, and as such, we should focus on oblique rotations rather than orthogonal. We will discuss these options in more detail later in Chapter 2.

Step 5: Interpreting results. Remember that the goal of exploratory factor analysis is to explore whether your data fits a model that makes sense. Ideally, you have a conceptual or theoretical framework for the analysis- a theory or body of literature guiding the development of an instrument, for example. Even if you do not, the results should be sensible in some way. You should be able to construct a simple narrative describing how each factor, and its constituent variables, makes sense and is easily labeled. It is easy to get EFA to produce results. It is much harder to get sensible results.

Note also that EFA is an *exploratory* technique. As such, it should not be used, as many researchers do, in an attempt to *confirm* hypotheses or test competing models. That is what *confirmatory factor analysis* is for. It is a misapplication of EFA to use it in this way, and we need to be careful to avoid confirmatory language when describing the results of an exploratory factor analysis.

If your results do not make sense, it might be useful to return to an earlier step. Perhaps if you extract a different number of factors, the factors or solution will make sense. This is why it is an exploratory technique.

Step 6: Replication of results. One of the hallmarks of science is replicability, or the ability for other individuals, using the same materials or methods, to come to the same conclusions. We have not historically placed much emphasis on replication in the social sciences, but we should. As you will see in subsequent chapters, EFA is a slippery technique, and the results are often not clear. Even clear results often do not replicate exactly, even within an extremely similar data set. Thus, in my mind, this step is critical. If the results of your analysis do not replicate, (or do not reflect the true nature of the variables in the "real world,") then why should anyone else care about them? Providing evidence that your factor structure is likely to replicate (either through another EFA or through CFA) makes your findings stronger and more relevant. In Chapter 5, we will explore a "traditional" method of replication[4] (similar to cross-

---

[4] I put that in quotations as most researchers reporting results from an EFA fail to do any

validation in regression models). In Chapter 6 we will play with the notion of applying a less traditional but perhaps more useful analysis using bootstrap analysis. Confirmatory factor analysis is outside the scope of this book, but is perhaps an even better method of replication.

Chapter 1 Summary

The goal of this book is to be a friendly, practical, applied overview of best practices in EFA. In the coming chapters, we will explore various aspects of EFA, and the best practices at each step. We will skip the matrix algebra and equations as much as possible. If you are seeking a more mathematical tome on the subject, there are many good ones already available.

You can refer to my website (http://jwosborne.com) for updates and errata (hopefully minimal) as well as data sets used in this book, syntax examples, how-to guides, answer keys to selected exercises, and other things as I think of them. . Please do not hesitate to contact me about this or any of my other books at jasonwosborne@gmail.com.

---

replication at all.

2 EXTRACTION AND ROTATION

Extraction is the general term for the process of reducing the number of dimensions being analyzed from the number of variables in the data set (and matrix of associations) into a smaller number of factors. Depending on the particular type of extraction, the type of association matrix being analyzed can be a matrix of simple correlations (the most commonly used type of association, and the default type of association analyzed in most statistical packages), but it could also be a matrix of covariances (which is more commonly analyzed in confirmatory factor analysis). Correlations are most commonly used in EFA as they are only influenced by the magnitude of the association of the two variables, whereas covariances are influenced by association, as well as the variance of each of the two variables in question (Thompson, 2004).

Extraction of factors proceeds generally (again, depending on the specific details of the extraction method chosen) by first extracting the strongest factor that accounts for the most variance, and then progressively extracting successive factors that account for the most remaining variance.

Choosing an extraction technique

Principal axes factor (PAF) extraction begins with initial estimates of communality coefficients, which can be obtained either from a principal components analysis or as a multiple regression equation predicting each variable from all other variables (multiple R^2) to provide starting values. Communality coefficients can be considered lower-bound estimates of score reliability to provide starting values (initial extraction), and also are the amount of variance accounted for in that variable by all other common factors combined (Cattell, 1965). Following this initial estimation, communality estimates are used to replace the diagonal elements of the correlation matrix (where PCA uses 1.0 on the diagonal elements signifying the expectation of perfect reliability of measurement). This substitution is important, as it acknowledges the realistic expectation of imperfect measurement. A new set of factors and communality coefficients are then estimated and the process is repeated iteratively until the communality coefficients stabilize- or change less than a pre-determined threshold.

If an EFA analysis fails to "converge" that means that these coefficients failed to stabilize and continued changing dramatically. This is most commonly due to inappropriately small sample sizes, but increasing the default number of iterations (often 25, as in SPSS) can help in certain cases. This extraction technique tends to be favored when multivariate normality of the variables is not a tenable assumption (Fabrigar et al., 1999).

Maximum Likelihood (ML) extraction[5] is another iterative process (used in logistic regression, confirmatory factor analysis, structural equation modeling, etc.) that seeks to extract factors and parameters that optimally reproduce the population correlation (or covariance) matrix. Starting with an assumption that individual variables are normally distributed (leading to multivariate normal distributions. If a certain number of factors are extracted to account for inter-relationships between the observed variables, then that information can be used to reconstruct a reproduced correlation matrix. The parameters chosen are tweaked iteratively in order to maximize the likelihood of reproducing the population correlation matrix- or to minimize the difference between the reproduced and population matrices. This technique is particularly sensitive to quirks in the data, particularly in "small" samples, so if the assumptions of normality are not tenable, this is probably not a good extraction technique (Fabrigar et al., 1999; Nunnally & Bernstein, 1994).

Unweighted Least Squares (ULS) and Generalized Least Squares (GLS) extraction both use variations on the same process of Maximum Likelihood extraction. ULS is said to be more robust to non-normal data (as we will see in the second example to come), whereas GLS weights variables with higher correlations more heavily, which can contribute to sensitivity to problematic data.

Alpha extraction seeks to maximize the Cronbach's alpha estimate of the reliability of a factor. The difference between alpha extraction and other extraction techniques is the goal of the generalization. ML and other similar extraction techniques seek to generalize from a sample to a population of individuals, whereas alpha extraction seeks to generalize to a *population of measures*. One downside to alpha extraction is that these properties are lost if rotation is used (Nunnally & Bernstein, 1994), applying only to the initial rotation. As we will see in the section on rotation, unrotated results are often not easily interpreted, leaving this extraction technique as a potentially confusing procedure where researchers may think they are getting something they are not if they rotate the results of the alpha extraction.

Initial communalities vs. extracted communalities. From the discussion above, we can expect to see differences in communalities and eigenvalues across extraction techniques, but one curious aspect of SPSS is that the initial communalities will always be the same.

---

[5] Interested readers can get much more information on this and other technical details of extraction in Nunnally & Bernstein (1994), particularly Chapter 11. For my purposes, I am attempting to provide a more applied overview, and so will eschew most of the technical details.

In PCA, the initial communalities are always 1.00 for all variables. In EFA, initial communalities are estimates of the variance in each variable that will be accounted for by all factors, arrived at either by PCA analysis or as some form of multiple regression type analysis. For example, initial communalities are estimated by a multiple regression equation predicting each variable from all other variables (multiple R^2). Extracted communalities are calculated as the variance accounted for in each variable by all extracted factors. Looking at a table of factor loadings, with variables as the rows and factor loadings in columns, communalities are row statistics. Squaring and summing each factor loading for a variable should equal the extracted communality (within reasonable rounding error).[6]

Initial eigenvalues vs. extracted eigenvalues vs. rotated eigenvalues. Eigenvalues are column statistics—again imagining a table of factor loadings, if you square each factor loading and sum them all within a column, you should get the eigenvalue for that factor (again within rounding error). Thus, eigenvalues are higher when there are at least some variables with high factor loadings, and lower when there are mostly low loadings. You will notice that eigenvalues (and communalities) change from initial statistics (which are estimates and should be identical regardless of extraction method as long as the extraction method is a true factor analysis extraction, not PCA) to extraction, which will vary depending on the mathematics of the extraction. The cumulative percent variance accounted for by the extracted factors will not change once we rotate the solution (discussed below in sections to come) but the distribution of that percent variance will change as the factor loadings change with rotation. Thus, if the extracted eigenvalues account for a cumulative 45% of the variance overall, once rotation occurs, the cumulative variance accounted for will still be 45%, but that 45% might be redistributed across factors. This will become clearer in a little bit, hopefully, as we look at some example data.

Three pedagogical examples

To illustrate the points in this section of the chapter, I have three example data sets that will also be available on the book web site (http://jwosborne.com).

Example 1: Engineering items. The first example is from a study on engineering majors at an eastern university. There were many scales and questionnaires administered, but we will concern ourselves with two: engineering problem solving and interest in engineering. These two scales should be at least minimally correlated, and give a relatively clear factor structure. The sample was composed of 372 undergraduate students. The items from the relevant subscales are listed below:

Problem solving items. How well did you feel prepared for:[7]

---

[6] where an orthogonal rotation is used. Oblique rotations are slightly more complicated if doing the calculation by hand, as the factors are correlated. Communalities in this case are the sum of each variable's pattern loading multiplied by the structure loading. Statistical software handles these complexities for us

1. Defining what the problem really is
2. Searching for and collecting information needed to solve the problem
3. Thinking up potential solutions to the problem
4. Detailing how to implement the solution to the problem
5. Assessing and passing judgment on a possible or planned solution to the problem
6. Comparing and contrasting two solutions to the problem on a particular dimension such as cost
7. Selecting one idea or solution to the problem from among those considered
8. Communicating elements of the solution in sketches, diagrams, lists, and written or oral reports

Interest in engineering:[8]

1. I find many topics in engineering to be interesting
2. Solving engineering problems is interesting to me
3. Engineering fascinates me
4. I am interested in solving engineering problems
5. Learning new topics in engineering is interesting to me
6. I find engineering intellectually stimulating

Example 2: Marsh Self-Description Questionnaire (SDQ). The second example includes three subscales from Marsh's Self-Description Questionnaire (SDQ; see e.g., Marsh, 1994). Data for this example is from 15,661 students in 10th grade with complete data on all items below who participated in the National Education Longitudinal Study of 1988 (NELS88, available from IES/NCES: http://nces.ed.gov/surveys/nels88/). Items from this scale include:

Parents:
F1S63A   (Par1)   My parents treat me fairly
F1S63F   (Par2)   I do not like my parents very much
F1S63I   (Par3)   I get along well with my parents
F1S63M (Par4)   My parents are usually unhappy or disappointed with what I do
F1S63U   (Par5)   My parents understand me

English
F1S63B   (Eng1)   I learn things quickly in English classes
F1S63E   (Eng2)   English is one of my best subjects
F1S63G   (Eng3)   I get good marks in English
F1S63N   (Eng4)   I'm hopeless in English classes

---

[7] Assessed on a seven point Likert type scale anchored by "did not prepare me at all" to "prepared me a lot"
[8] Assessed on a seven point Likert type scale anchored by "strongly disagree" and "strongly agree"

Mathematics
F1S63D   (Math1)   Mathematics is one of my best subjects
F1S63J   (Math2)   I have always done well in mathematics
F1S63Q   (Math3)   I get good marks in mathematics
F1S63S   (Math4)   I do badly in tests of mathematics

Example 3: Geriatric Depression Scale. The third example is data on the Geriatric Depression Scale (GDS), a 30-item scale that does not seem to have a clear factor structure. These data are on N=479 older adults from the Long Beach Longitudinal Survey.[9] All items are scored either 0 (non-depressive answer) or 1 (depressive answer),[10] and it was originally designed to have 5 subscales. However, there has been considerable debate in the literature as to the true factor structure. I often use these data in my classes as it reveals the art (and frustration, occasionally) of exploratory factor analysis. The items to the GDS are as follows:

1. Are you basically satisfied with your life?
2. Have you dropped many of your activities and interests?
3. Do you feel that your life is empty?
4. Do you often get bored?
5. Are you hopeful about the future?
6. Are you bothered by thoughts you can't get out of your head?
7. Are you in good spirits most of the time?
8. Are you afraid that something bad is going to happen to you?
9. Do you feel happy most of the time?
10. Do you often feel helpless?
11. Do you often get restless and fidgety?
12. Do you prefer to stay at home, rather than going out and doing new things?
13. Do you frequently worry about the future?
14. Do you feel you have more problems with memory than most?
15. Do you think it is wonderful to be alive now?
16. Do you often feel downhearted and blue?
17. Do you feel pretty worthless the way you are now?
18. Do you worry a lot about the past?
19. Do you find life very exciting?
20. Is it hard for you to get started on new projects?
21. Do you feel full of energy?
22. Do you feel that your situation is hopeless?
23. Do you think that most people are better off than you are?
24. Do you frequently get upset over little things?
25. Do you frequently feel like crying?
26. Do you have trouble concentrating?

---

[9] Zelinski, Elizabeth, and Robert Kennison. Long Beach Longitudinal Study. ICPSR26561-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2011-06-17. http://doi.org/10.3886/ICPSR26561.v2
[10] And for this data set, all items were recoded so that 0 always is a non-repressive answer and 1 is the depressive answer even when the items were initially reversed.

27. Do you enjoy getting up in the morning?
28. Do you prefer to avoid social gatherings?
29. Is it easy for you to make decisions?
30. Is your mind as clear as it used to be?

Does extraction method matter?

It is common wisdom that extraction techniques tend to yield similar results. Let us examine this assertion to see if it holds across our samples.

Example 1: Analysis of the engineering data. These data usually give a clear two-factor solution. The goal of this analysis is to examine whether we get substantially different results as a function of extraction method. As you can see in Table 2.1a, the communalities for the various items are relatively stable despite the relatively small sample size, rarely varying by more than 0.10 across all extraction methods for a particular item. In Table 2.1b, you can see that eigenvalues from ML and PAF extraction also produced similar results (as did most of the other extraction techniques).

Table 2.1a
Communality estimates for the engineering data across different extraction techniques.

Variable:	Initial	ML	PAF	ULS	GLS	Alpha
EngProbSolv1	.742	.712	.728	.728	.809	.733
EngProbSolv2	.695	.663	.669	.669	.757	.669
EngProbSolv3	.752	.765	.768	.768	.794	.769
EngProbSolv4	.792	.810	.810	.810	.843	.810
EngProbSolv5	.790	.807	.799	.799	.832	.796
EngProbSolv6	.766	.774	.768	.768	.813	.767
EngProbSolv7	.786	.778	.775	.775	.845	.774
EngProbSolv8	.666	.674	.671	.671	.705	.669
INTERESTeng1	.674	.666	.669	.668	.725	.668
INTERESTeng2	.802	.834	.833	.833	.846	.834
INTERESTeng3	.816	.847	.840	.840	.864	.839
INTERESTeng4	.806	.831	.817	.817	.853	.813
INTERESTeng5	.781	.781	.800	.800	.842	.805
INTERESTeng6	.739	.750	.752	.752	.784	.751

Comparison of the two recommended extraction techniques, ML and PAF, produced similar results, suggesting that when basic assumptions are met and factor structure is clear, the extraction method might not matter much.

Table 2.1b.
Eigenvalues extracted for the engineering data across different extraction techniques

Factor:	Initial	ML	PAF	ULS	GLS	Alpha
1	7.653	7.359	7.417	7.417	7.439	7.415
2	3.505	3.335	3.282	3.282	3.341	3.285
3	.457					
4	.360					
5	.315					

Note: factors 6-14 suppressed from initial extraction. Only two factors extracted.

Table 2.2a
Communality estimates for the SDQ data across different extraction techniques.

Variable:	Initial	ML	PAF	ULS	GLS	Alpha
Eng1	.537	.619	.623	.623	.631	.621
Eng2	.581	.676	.664	.664	.718	.648
Eng3	.608	.722	.723	.724	.738	.722
Eng4	.447	.403	.413	.413	.561	.425
Math1	.704	.790	.792	.792	.812	.794
Math2	.674	.751	.737	.737	.753	.721
Math3	.700	.783	.799	.800	.792	.816
Math4	.393	.372	.371	.371	.453	.374
Par1	.455	.526	.510	.510	.545	.496
Par2	.406	.434	.450	.450	.500	.458
Par3	.572	.695	.678	.678	.718	.668
Par4	.408	.392	.421	.421	.501	.442
Par5	.477	.557	.539	.539	.575	.525

Table 2.2b.
Eigenvalues extracted for the SDQ data across different extraction techniques

Factor:	Initial	ML	PAF	ULS	GLS	Alpha
1	4.082	3.399	3.689	3.689	3.450	3.622
2	2.555	2.446	2.226	2.226	2.456	2.258
3	2.208	1.874	1.804	1.804	1.916	1.829
4	.908					
5	.518					
6	.487					

Note: factors 7-13 suppressed from initial extraction. Only two factors extracted.

Example 2: Analysis of the Self-Description Questionnaire data. As in the previous analysis, you can see that analyses of the SDQ by various extraction methods produce relatively similar results regardless of the extraction method (presented in Table 2.2a). The communalities extracted were similar, and the eigenvalues were also similar (presented in Table 2.2b).

Example 3: GDS data. The goal of the third analysis is to compare the results of

Table 2.3a
Comparison of communalities across extraction methods

Variable:	Initial	ML[1]	PAF	ULS	GLS[1]	Alpha
GDS01	.518	.880	.689	.685	.988	.553
GDS02	.297	.346	.366	.366	.383	.367
GDS03	.513	.561	.579	.580	.567	.558
GDS04	.408	.612	.576	.576	.683	.550
GDS05	.400	.424	.396	.394	.555	.398
GDS06	.369	.450	.447	.446	.470	.447
GDS07	.451	.543	.522	.520	.572	.436
GDS08	.272	.276	.329	.331	.392	.391
GDS09	.559	.689	.672	.671	.771	.629
GDS10	.410	.416	.406	.406	.487	.397
GDS11	.310	.364	.372	.372	.402	.349
GDS12	.320	.988	.718	.813	1.000	.659
GDS13	.278	.428	.389	.382	.486	.314
GDS14	.286	.406	.451	.458	.454	.489
GDS15	.384	.409	.430	.430	.493	.470
GDS16	.534	.564	.567	.567	.636	.561
GDS17	.500	.553	.548	.545	.641	.531
GDS18	.290	.264	.281	.283	.420	.314
GDS19	.396	.422	.420	.419	.480	.411
GDS20	.336	.355	.387	.388	.465	.462
GDS21	.346	.417	.435	.433	.461	.433
GDS22	.413	.471	.491	.491	.566	.514
GDS23	.254	.254	.252	.252	.336	.264
GDS24	.260	.280	.282	.283	.349	.311
GDS25	.442	.451	.473	.474	.574	.482
GDS26	.375	.445	.437	.435	.547	.425
GDS27	.211	.214	.239	.240	.275	.260
GDS28	.300	.310	.346	.329	.328	.336
GDS29	.195	.162	.168	.168	.279	.219
GDS30	.277	.368	.362	.363	.397	.380

1. Produced a warning about communality estimates greater than 1 were encountered during iterations.

various extraction techniques on data with less clarity of structure. Because these data are binary (0, 1 values only) it is likely that they do not meet the assumption of multivariate normality. If one takes the advice above seriously, PAF would probably be the ideal extraction technique given the non-normal data. As you can see in Table 2.3a, there are substantial differences in some communalities extracted for several variables in the scale. Comparing ML and PAF, it is clear that the recommendation to use PAF when data are not multivariate normal should be seriously considered.

There are some items in Table 2.3a that exhibited substantial discrepancies between ML and PAF (for example, see GDS01 or GDS12). During ML and GLS extraction, there was a warning that some of the iterations observed estimated communalities greater than 1.00 (an impossible number). This is generally a sign of a serious issue with the analysis. In this case, it is likely due to violation of the assumption of normality.

The eigenvalues extracted also vary dramatically across extraction methods, as you can see in Table 2.3b. Once again, ML and generalized least squares produce the most unexpected results, while PAF results are probably most reliable. Surprisingly, ML and GLS produced smaller eigenvalues for the first factor than the second factor, which is unusual (and also a sign of potential problems).

Table 2.3 b
Comparison of extracted eigenvalues for different extraction techniques

Factor:	Initial	ML	PAF	ULS	GLS	Alpha
1	7.858	2.431	7.324	7.325	2.090	7.296
2	2.079	5.642	1.560	1.567	3.792	1.539
3	1.702	1.407	1.116	1.125	3.464	1.115
4	1.318	1.028	.773	.794	1.322	.744
5	1.186	.975	.683	.702	1.075	.635
6	1.137	.754	.620	.626	.817	.649
7	1.083	.623	.535	.538	.706	.512
8	1.020	.457	.420	.421	.532	.419
9	.929					
10[1]	.907					

1.   Factors 11-30 suppressed in this example.

Summary: Does extraction method matter? While there are many options for extraction in most statistical computing packages, there is consensus that ML is the preferred choices for when data exhibit multivariate normality and PAF for when that assumption is violated. Other extraction techniques (GLS, in particular) seem to be vulnerable to violations of this assumption, and do not seem to provide any substantial benefit. Thus, the general recommendation to use either ML or PAF seems sensible.

Deciding how many factors should be extracted and retained

Scholars have been arguing about this issue for the better part of a century. There are many who will argue passionately that one method is superior to another, and some do seem to be more defensible than others. However, it seems to me that much of the argument comes from a high-stakes mentality where researchers are attempting to *confirm* one factor structure as superior to another. Let us again repeat our mantra for this book: EFA is exploratory and should be considered a low-stakes process.

There are many guidelines for how to decide the number of factors to extract from an analysis. After all, your statistical computing software will extract as many factors as there are variables in the analysis, and since our goal is dimension reduction, we then have to decide how many of those extracted factors to retain for further analysis. So what decision rules are best?

Theory. I am a big proponent of theory-driven analysis.[11] Researchers often perform an EFA because someone designed an instrument to measure particular constructs or factors. If the theoretical framework for the instrument is sound, we should start with the expectation that we should see that structure in the data. Instruments are rarely perfect (especially the first time it is examined), and theoretical expectations are not always supported. But unless one is on a fishing expedition in a data set with no *a priori* ideas about how the analysis should turn out,[12] this is as good a place as any to start. Regardless, the result of an EFA must be a sensible factor structure that is easily understood, whether that final structure matches the initial theoretical framework or not. The basic function of EFA, in my mind, is to make meaning of data.

The Kaiser Criterion. The default in most statistical software packages is use the Kaiser criterion (Kaiser, 1960, 1970), which proposed that an eigenvalue greater than 1.0 is a good lower bound for expecting a factor to be meaningful. This is because an eigenvalue represents the sum of the squared factor loadings in a column, and to get a sum of 1.0 or more, one must have rather large factor loadings to square and sum (e.g., four loadings of at least 0.50 each, three loadings of at least 0.60 each). But this criterion gets less impressive as more items are analyzed. It is easy to get many unimportant factors exceeding this criterion if you analyze 100 items in an analysis.

Despite the consensus in the literature that this probably the least accurate method for selecting the number of factors to retain (Velicer, Eaton, & Fava, 2000; see also Costello & Osborne, 2005), it is usually implemented as the default selection criteria in statistical software. Prior to the wide availability of powerful computing, this was a simple (and not unreasonable) decision rule. Toward the later part of the 20th century, researchers suggested that combining this criterion with examination of the scree plot is better (Cattell, 1966).

---

[11] Not just in EFA but in all research. If we are not purposeful and thoughtful, why are we doing this?

[12] To be clear, I strongly discourage this sort of analysis. Concepts and theories should always guide what we do when exploring data. Except when they don't.

Scree plot. The scree test involves examining the graph of the eigenvalues (available via every software package) and looking for the natural bend or "elbow" in the data where the slope of the curve changes (flattens) markedly. The number of data points above the "elbow" (i.e., not including the point at which the break occurs) is usually considered a good estimate of the ideal number of factors to retain. Although the scree plot itself is not considered sufficient to determine how many factors should be extracted (Velicer et al., 2000), many suggest that researchers examine solutions extracting the number of factors ranging from one to two factors above the elbow to one or two below. As this is an exploratory technique, one should be encouraged to explore. Some scree plots do not have one clear bend. Some have multiple possible points of inflection, and some have no clear inflection point (for a good example of this, see the SDQ example, below). Combining theory, the Kaiser criterion, and examination of the scree plot is usually a good basis for deciding the number of factors to extract in an exploratory factor analysis.

Parallel analysis was proposed by Horn (1965). It is not widely included in common statistical computing packages, and thus is not widely used. However, it is considered advantageous over the more classic approaches (although we will see in examples below that it is not always better; c.f., Velicer et al., 2000). Parallel analysis involves generating random uncorrelated data, and comparing eigenvalues from the EFA to those eigenvalues from those random data. Using this process, only factors with eigenvalues that are *significantly* above the mean (or preferably, the 95th percentile) of those random eigenvalues should be retained. Several prominent authors and journals have endorsed this as the most robust and accurate process for determining the number of factors to extract (Ledesma & Valero-Mora, 2007; Velicer et al., 2000). The problem is that common statistical computing packages, such as SPSS and SAS, have not incorporated this as an option for researchers.

Minimum Average Partial (MAP) criteria was proposed by Velicer (1976) as another more modern methodology for determining the number of factors to extract in the context of PCA. This procedure involves partialing out common variance as each successive component is created; a familiar concept to those steeped in the traditions of multiple regression. As each successive component is partialed out, common variance will decrease to a minimum. At that point, unique variance is all that remains. Velicer argued that minimum point should be considered the criterion for the number of factors to extract (Velicer et al., 2000). MAP has been considered superior to the "classic" criteria, and probably is superior to parallel analysis, although neither is perfect, and all must be used in the context of a search for conceptually interpretable factors.

Using MAP and parallel analysis. One barrier to researchers using MAP and parallel analysis is that these procedures are not widely implemented in statistical software packages. For those wanting to experiment with them, freely-downloadable FACTOR software.[13] For users of SPSS, SAS, or MATLAB, O'Connor (2000) has

---

[13] Lorenzo-Seva and Ferrando (2006) have made their software freely available from their website: http://psico.fcep.urv.es/utilitats/factor/

provided simple and efficient programs for performing these less common but more ideal procedures. These can currently be downloaded from https://people.ok.ubc.ca/brioconn/boconnor.html . I will also make copies available on the website for this book for ease of use.

While I can understand the value of parallel analysis or MAP criteria for deciding how many factors to extract, we have to remember our mantra: EFA is an exploratory technique. No criterion is perfect, and unless you are mis-using EFA in a confirmatory fashion, it seems to me that worrying over a slightly better extraction criterion might be missing the point. The point is to get a reasonable model within a representative sample (that is sufficiently large to ensure a reasonable solution), and then to move into inferential statistics available in confirmatory factor analysis. EFA is merely a first stopping point on the journey, and researchers who forget this miss the point of the process. Thus, use parallel analysis or MAP criteria, along with theory (and any of the classic criteria that suits you and is defensible). The goal of creating theory-driven, conceptually understandable solutions needs to prevail. And never forget that your journey is not done until you confirm the results of the EFA in the context of CFA.

Example 1: Engineering data

The factor structure for these data are expected to be clear: two factors, one reflecting engineering problem solving and one reflecting interest in engineering. Theory is our ally in this case as these scales were theory-driven. If we look at the eigenvalues from Table 2.1b, we see that regardless of the extraction method, there are two factors with strong eigenvalues over 3.00, and no other factors with eigenvalues above 0.50. Thus, according to the Kaiser criterion we should extract two factors. Next, the Scree plot (Figure 2.1) shows a pronounced inflection point at factor #3, indicating that we should consider 2 factor solution (and explore 2, 3, and 4 factor solutions, perhaps, if we did not have strong theory behind the two-factor solution). Three classic indicators suggest a two-factor solution is ideal for these data (until we explore the data in the context of confirmatory analyses).

Examine the factor loadings (since we have not talked about rotation yet, we will examine unrotated factor loadings) in Table 2.4, also plotted in Figure 2.2. As you can see from the factor loading plot in Figure 2.2, these data are clustered into to clear and distinct factors. In my opinion, this is about as good as it gets in exploratory factor analysis at this stage of the analysis, and examining MAP criteria or performing parallel analysis to determine how many factors to extract would not result in any more useful information. However, for the sake of providing an example, these same data were subjected to first parallel and then MAP analysis.[14]

---

[14] The syntax and data for this example are available on the book web site at http://jwosborne.com. Syntax for SPSS, as mentioned above, was adapted from (O'connor, 2000).
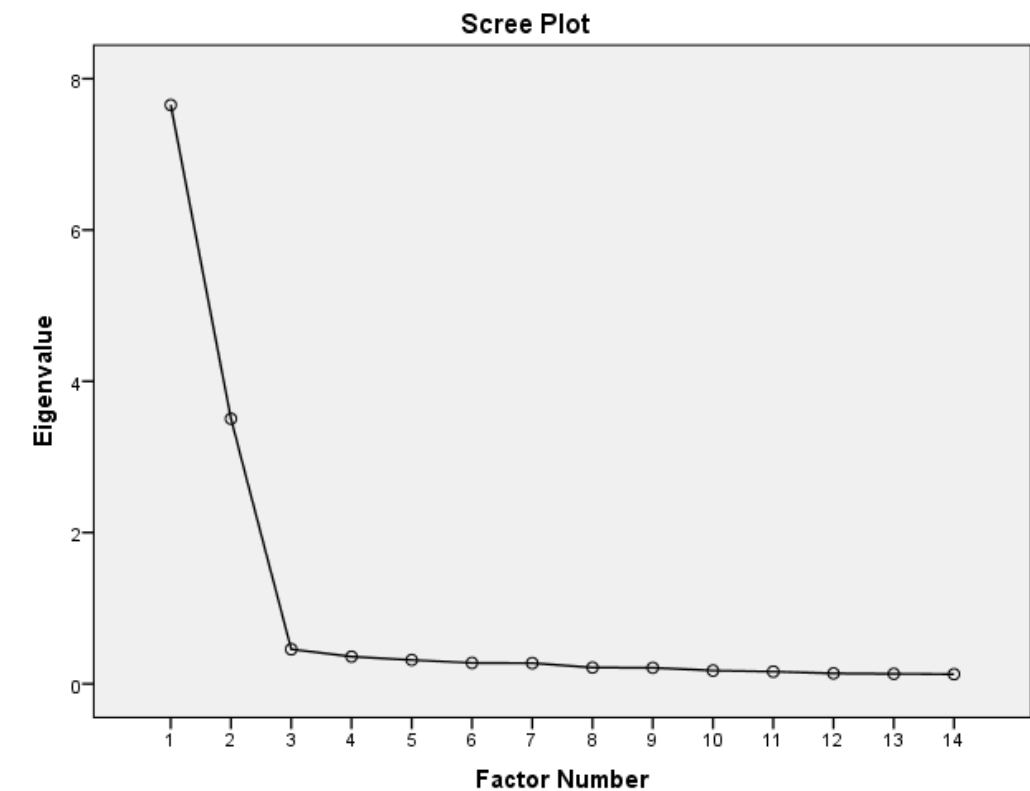
Figure 2.1: Scree plot from engineering data, ML extraction

Table 2.4
Unrotated Factor Matrix for Engineering Data

Variable:	Factor	
	1	**2**
EngProbSolv1	.706	.463
EngProbSolv2	.649	.492
EngProbSolv3	.731	.481
EngProbSolv4	.743	.508
EngProbSolv5	.764	.473
EngProbSolv6	.748	.464
EngProbSolv7	.754	.458
EngProbSolv8	.723	.389
INTERESTeng1	.688	-.439
INTERESTeng2	.730	-.549
INTERESTeng3	.740	-.546
INTERESTeng4	.738	-.535
INTERESTeng5	.719	-.514
INTERESTeng6	.711	-.495

Factor Plot
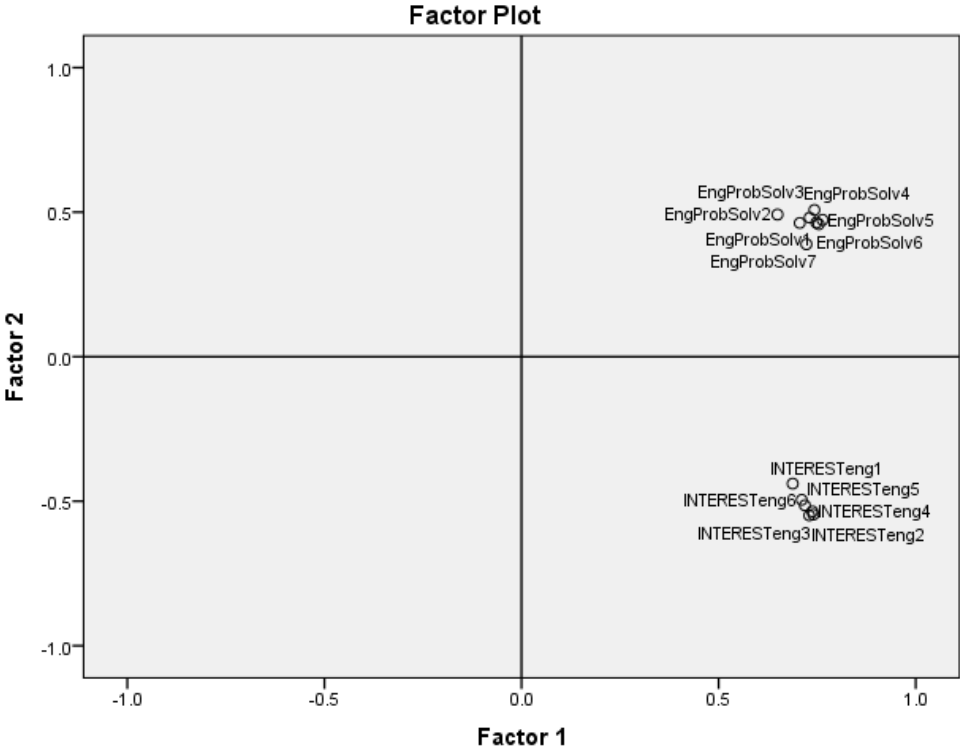
Figure 2.2: Factor loading plot Maximum Likelihood extraction

Parallel analysis. The parallel analysis first generated random data for mean and 95th percentile eigenvalues. The syntax provided on the book web site also produced the real-data eigenvalues from this sample for comparison. Using appropriate specifications (14 variables, appropriate sample size, etc.), the parallel analysis produced the following random data and real data eigenvalues, (which are plotted in Figure 2.3).

Recall that the goal is to select the number of factors to extract where the observed eigenvalues from the data are significantly higher than the random eigenvalues. In this case, parallel analysis indicates that two factors have eigenvalues exceeding that of random samples (95th percentiles used here, as recommended in the literature).

```
Random Data Eigenvalues
       Root          Means           95th%
    1.000000        .318540         .385296
    2.000000        .252631         .313547
    3.000000        .200020         .239804
    4.000000        .150248         .189150
    5.000000        .107605         .143019
    6.000000        .070462         .099675
    7.000000        .032913         .060748
```
<subsequent eigenvalues were negative and were deleted for space>

```
Raw Data Eigenvalues
       Root          Eigen.
    1.000000        7.411470
    2.000000        3.271016
    3.000000         .197160
    4.000000         .069918
    5.000000         .048940
    6.000000         .039340
    7.000000         .021963
```
<subsequent eigenvalues were negative and were deleted for space>
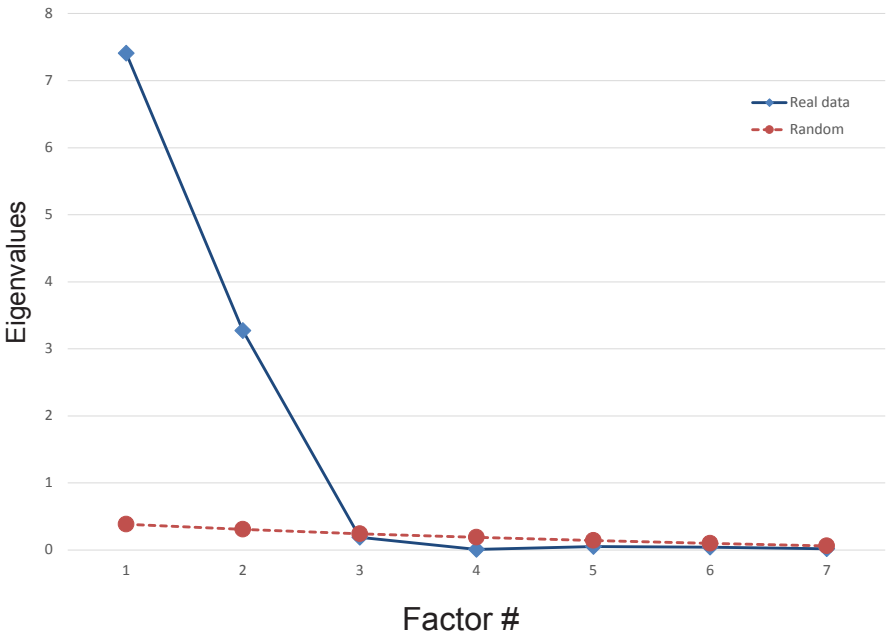
Figure 2.3. Plot of parallel analysis results comparing

MAP criteria. Finally, the MAP criteria were calculated for these data. Recall that using MAP criteria, we want to choose a number of factors where the inflection point where the graph of the average partial correlations hits a minimum.

```
Eigenvalues
       7.6534
       3.5048
        .4568
        .3597
```

<subsequent eigenvalues were negative and were deleted for space>

```
Average Partial Correlations
                  squared          power4
       .0000        .3167          .1645
      1.0000        .2458          .0743
      2.0000        .0236          .0018
      3.0000        .0295          .0034
      4.0000        .0436          .0088
      5.0000        .0580          .0119
      6.0000        .0724          .0229
      7.0000        .0926          .0436
```
<subsequent eigenvalues were negative and were deleted for space>

As you can see from the MAP test output and in Figure 2.4, the inflection point (minimum) on the plot of average partial correlations is at 2, and the output suggested two factors be extracted, which corresponds exactly with our theory and other extraction criteria for these data.
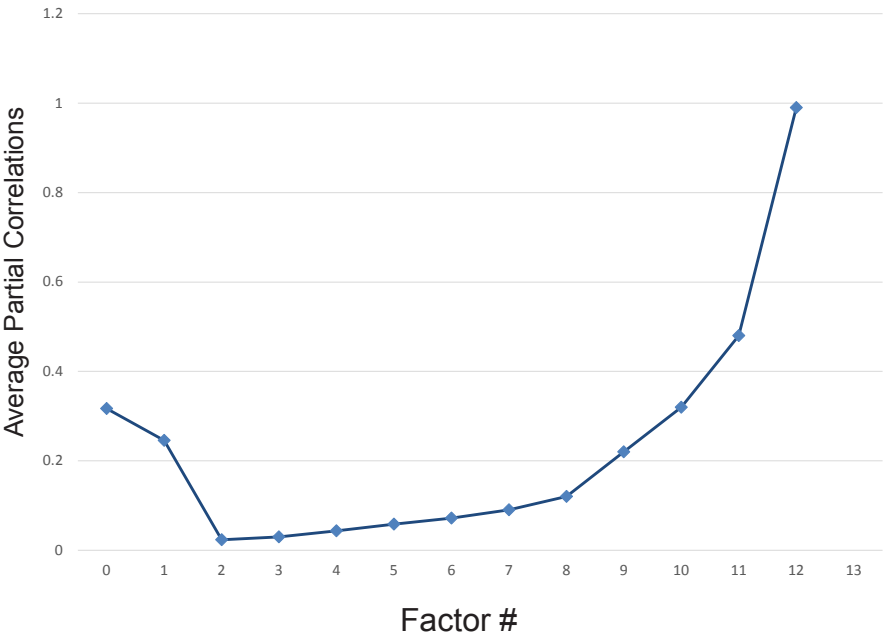
Figure 2.4: Plot of average partial correlations for MAP test

Summary of results for Example 1. Using several different methods of selecting the number of factors to extract, all methods pointed to the conclusion that two factors is the optimal number to extract. Given that the results make sense in the context of the theoretical model, we would extract two factors for rotation and interpretation.

Example 2: Self-Description Questionnaire data

Since the data from the SDQ seemed to be relatively well-behaved across the initial extraction comparisons, I will again use ML extraction and explore whether our expected three-factor solution is tenable. The first two criteria, theory and eigenvalues, all suggest a three-factor solution. In the case of these data, however, the scree plot (presented in Figure 2.5) is a bit less clear. Scree plots do not always have one clear elbow. In this case, it is possible to argue that any one of several points is the true "elbow"- including 2, 4, or 5. In this example, the scree plot is not terribly informative. Remember that this analysis had a very large sample, and thus might not be representative of analyses with smaller samples.
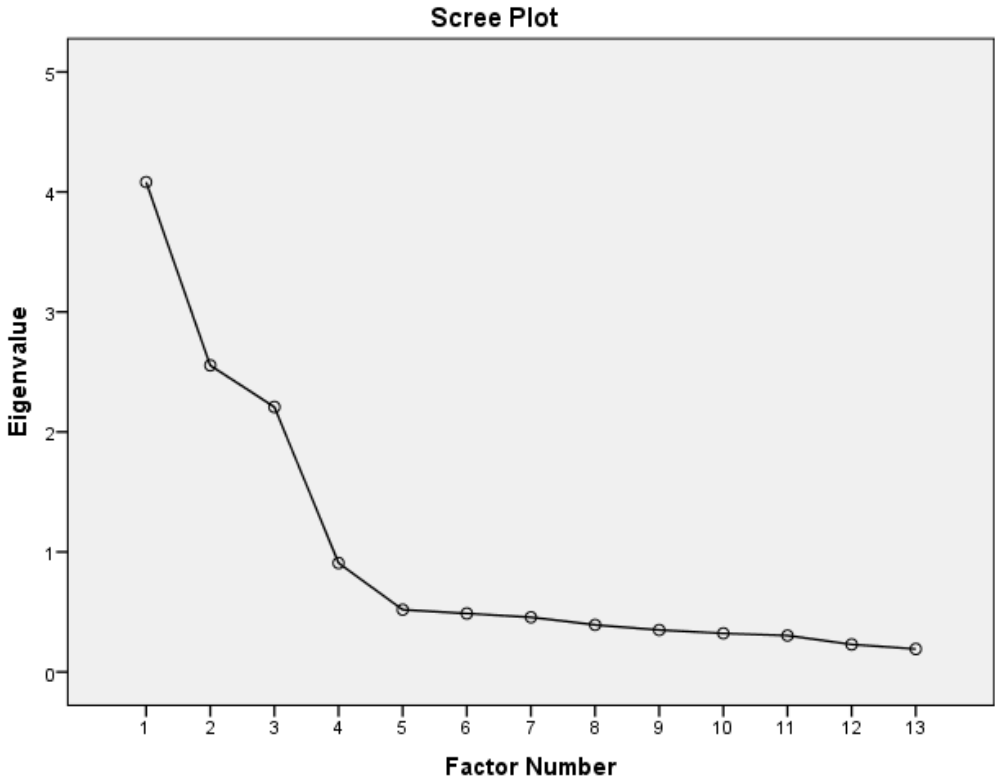
Figure 2.5: scree plot from ML extraction of SDQ data.

Parallel analysis. Because the sample size was so large, parallel analysis might not be as useful. The largest randomly generated eigenvalues (95th percentile) was 0.057. Thus, using the criteria for parallel analysis, one would recommend examining either three or four factors (depending on how "significantly" different the raw data eigenvalue should be).

```
Raw Data Eigenvalues
        Root        Eigen.
     1.000000      3.624965
     2.000000      2.158366
     3.000000      1.731491
     4.000000       .361588
     5.000000      -.021073
     6.000000      -.053630
     7.000000      -.062606
```
<subsequent eigenvalues were negative and were deleted for space>

```
Random Data Eigenvalues
        Root        Means        95th %
     1.000000      .047533       .056737
     2.000000      .036696       .044227
     3.000000      .028163       .035459
     4.000000      .020568       .025826
     5.000000      .013416       .018670
     6.000000      .006943       .011787
     7.000000      .000123       .005167
```
<subsequent eigenvalues were negative and were deleted for space>
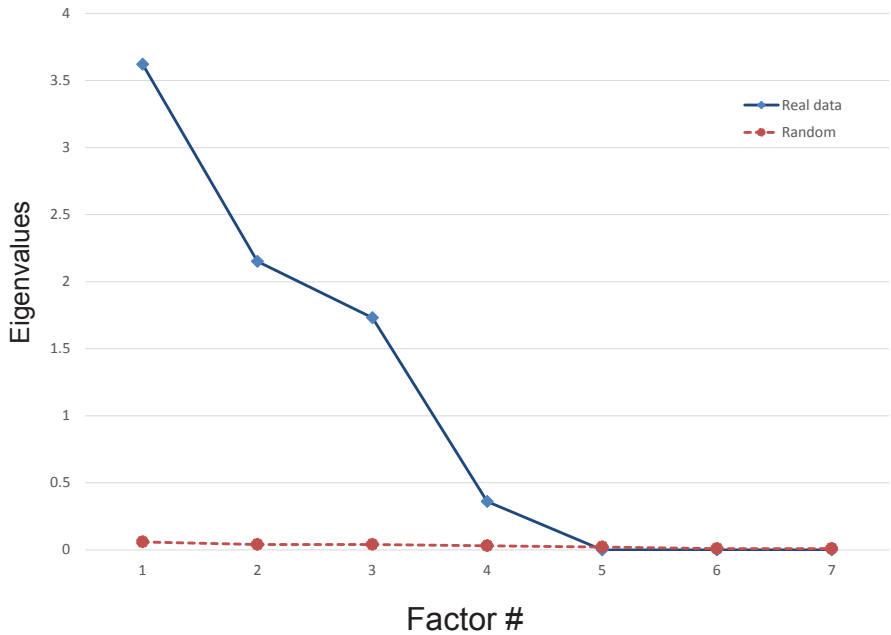
Figure 2.6: Parallel analysis of SDQ data

MAP criteria. The data from the MAP analysis seem to reinforce theory and other criteria, indicating that three factors is the right number to extract. As you can see in Figure 2.7, the minimum inflection point is at 3.

```
Average Partial Correlations
                    squared        power4
        .0000        .1100         .0375
       1.0000        .0993         .0178
       2.0000        .0854         .0124
       3.0000        .0349         .0037
       4.0000        .0401         .0046
       5.0000        .0606         .0127
       6.0000        .0884         .0257
```

<subsequent eigenvalues were negative and were deleted for space>

```
 The  Number  of  Components  According  to  the  Original
(1976) MAP Test is 3
 The  Number  of  Components  According  to  the  Revised
(2000) MAP Test is  3
```
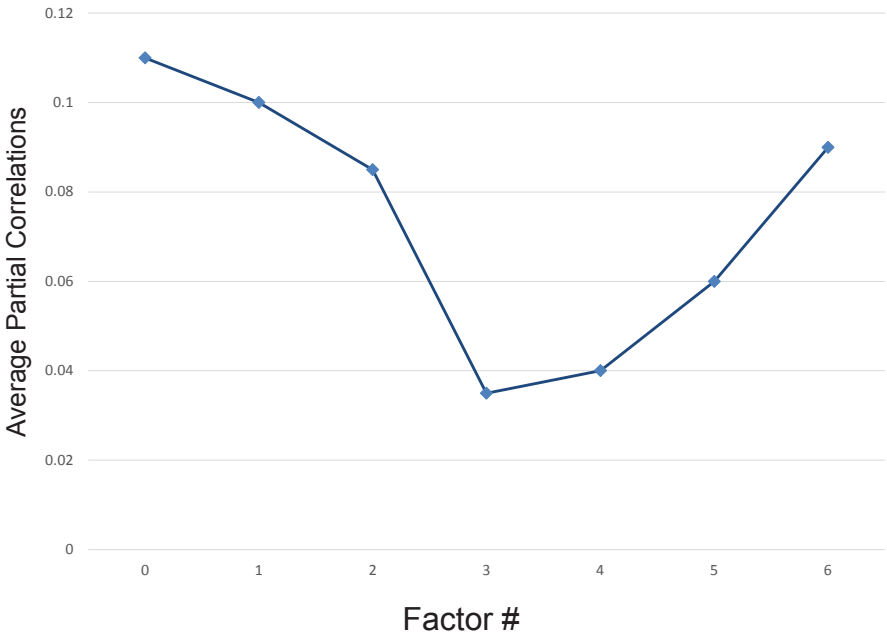

Figure 2.7: Plot of average partial correlations for MAP test

Summary of results for Example 2. As Velicer and others have argued, the MAP criteria appear to be, at least in this case, more congruent with theory and eigenvalues-which is reassuring. The parallel analysis criteria recommends extraction of four factors. The three factor model seems to be the best recommendation as it makes for a strong, interpretable model.

Example 3: Geriatric Depression Scale data

Referring back to Table 2.3b, this scale provided a very different counterpoint to the clear conceptually consistent results of the engineering and SDQ data. This scale was designed to have five subscales originally,[15] so theory would suggest that there are five factors. But as with many of our endeavors in the social sciences, this might not hold true once put to the test. For example, it is just as likely that all items will load as a single factor, or that a different number of factors will be ideal. The results of the PAF extraction from earlier in the chapter, there were eight factors that had eigenvalues greater than 1.0 (eigenvalue #8 was 1.02, and #9 was 0.93, leaving some ambiguity around whether this is a viable cutoff).
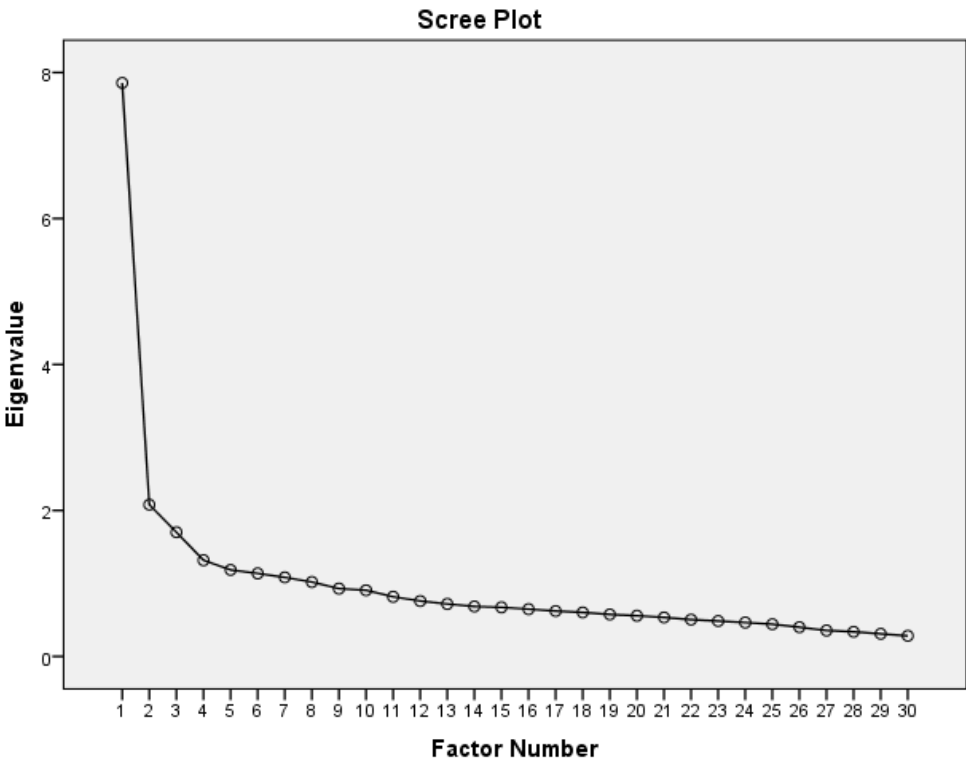
Figure 2.8: Scree plot of Geriatric Depression Scale data.

---

[15] Scale 1 was to be composed of item # 1, 6, 11, 16, 21, and 26; scale 2: #2, 7, 12, 17, 22, 27; and so forth.

Further, the scree plot, in Figure 2.8, seems to indicate that the first inflection point is at 2 factors, but it is also arguable that there is a second inflection point at the fourth factor. Thus, using traditional criteria, we would probably combine these results and test for a variety of configurations including 1, 3, 4, and 5 factor extraction. These results would be examined to see if the original theoretical framework made sense, or if any of the other factor structures seem to make sense. However, since we have parallel analysis and MAP criteria, let us examine those results before exploring these options.

Parallel analysis. Because of the large number of items in this scale, I will truncate the output to reasonable numbers of factors.

```
Raw Data Eigenvalues
        Root        Eigen.
    1.000000       7.257380
    2.000000       1.458959
    3.000000       1.044359
    4.000000        .674890
    5.000000        .548184
    6.000000        .484804
    7.000000        .434659
```
<subsequent eigenvalues were negative and were deleted for space>

```
Random Data Eigenvalues
        Root        Means        95th%
    1.000000       .569249      .644117
    2.000000       .495687      .540897
    3.000000       .446227      .491564
    4.000000       .397062      .436511
    5.000000       .354778      .393940
    6.000000       .318813      .357407
    7.000000       .281370      .319290
```
<subsequent eigenvalues were negative and were deleted for space>

The results of this parallel analysis poses an interesting dilemma, as the eigenvalues quickly drop below 1 in the raw data analysis, and quickly approach the random data eigenvalues, as you can see in the data above and Figure 2.9, below. However it is not until around the 8th factor that the lines meet, which incidentally is the same conclusion as the Kaiser criterion leads to (eight factors with eigenvalues greater than 1.0). This would suggest that we should extract many more factors than probably makes sense.
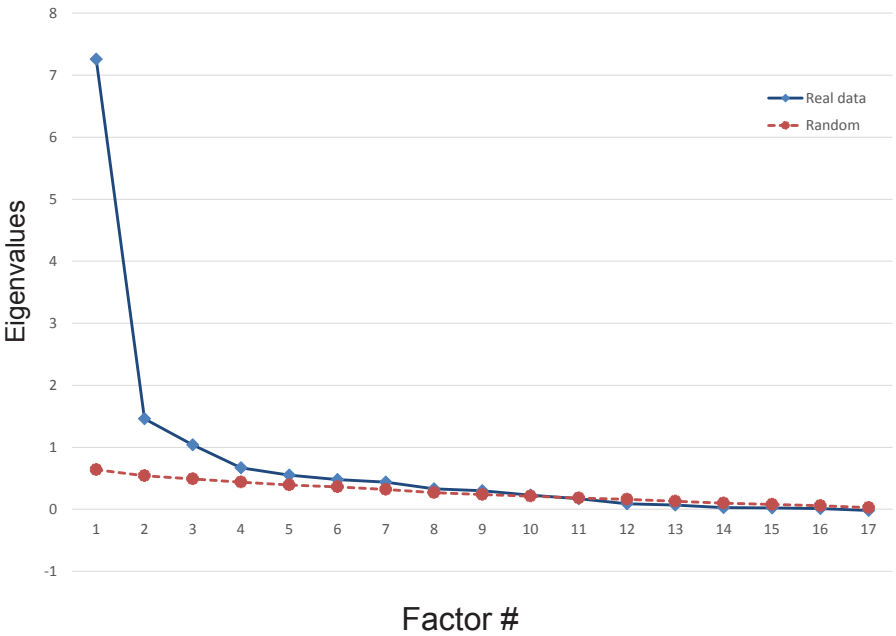
Figure 2.9. Parallel Analysis of GDS data.

MAP criteria. The MAP analysis was also interesting for these data, in that the MAP results recommend extraction of three factors, rather than the single strong factor of the theoretically-expected five factors. As you can see in the MAP plot (Figure 2.10), there might be a significant inflection point at the first factor. The true minimum is clearly at 3 factors, but the change between factors 2, 3, and 4 is so minimal as to be almost inconsequential. The third factor is only 0.0003 less than the APC for factor 2, and only 0.0008 less than factor 4. As you can see in Figure 2.10, one could argue that the only real inflection point is at 1.

```
Average Partial Correlations
                  squared         power4
        .0000        .0616         .0066
       1.0000        .0111         .0004
       2.0000        .0100         .0003
       3.0000        .0097         .0003
       4.0000        .0105         .0003
       5.0000        .0118         .0004
       6.0000        .0132         .0006
       7.0000        .0148         .0008
```
<subsequent eigenvalues were negative and were deleted for space>
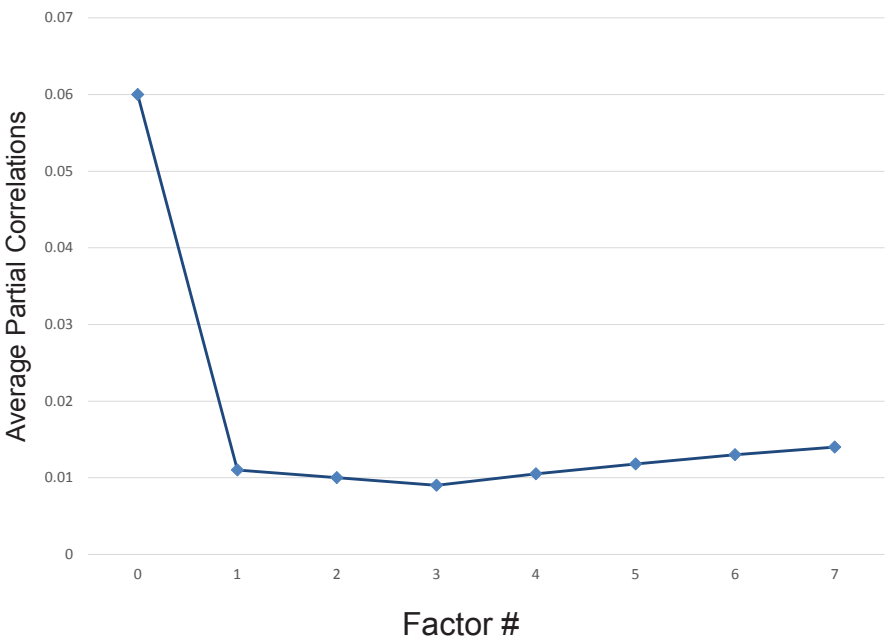
Figure 2.10: Plot of average partial correlations for MAP test

Summary of results for Example 3. This third example reinforces the fact that EFA is both an art and a quantitative science, and that good judgment of the researcher is critical when the data are not as clear or cooperative as one would hope. These data are not giving us a clear indication of how many factors to extract, and thus we need to *explore* several options for what is most sensible. When data are uncooperative in this way, replication becomes even more critical, as we will discuss in chapters to come.

Rotation in EFA

Unrotated results from a factor analysis – as presented above in example 1- is not easy to interpret, although the factor loading plot can help. Simply put, rotation was developed not long after factor analysis to help researchers clarify and simplify the results of a factor analysis. Indeed, early methods were subjective and graphical in nature (Thurstone, 1938) because the calculations were labor intensive. Later scholars attempted to make rotation less subjective or exploratory (e.g., Horst, 1941), leading to initial algorithms such as Quartimax (Carroll, 1953) and Varimax (Kaiser, 1958), which is currently the most common rotation (perhaps because it is the default in many statistical computing packages).[16]

---

[16] Note that rotation *does not* alter the basic aspects of the analysis, such as the amount of variance extracted from the items. Indeed, although eigenvalues might change as factor loadings are adjusted by rotation, the overall percentage of variance accounted for will

Quite simply, we use the term "rotation" because, historically and conceptually, the axes are being rotated so that the clusters of items fall as closely as possible to them.[17] As Thompson (2004) notes, the location of the axes are entirely arbitrary, and thus we can rotate the axes through space (like turning a dial) without fundamentally altering the nature of the results. However, we cannot move the location of any variable in the factor space.

Looking at Figure 2.11, for example, if you imagine rotating the axes so that they intersect the centroid of each cluster variables, you get the essence of rotation.
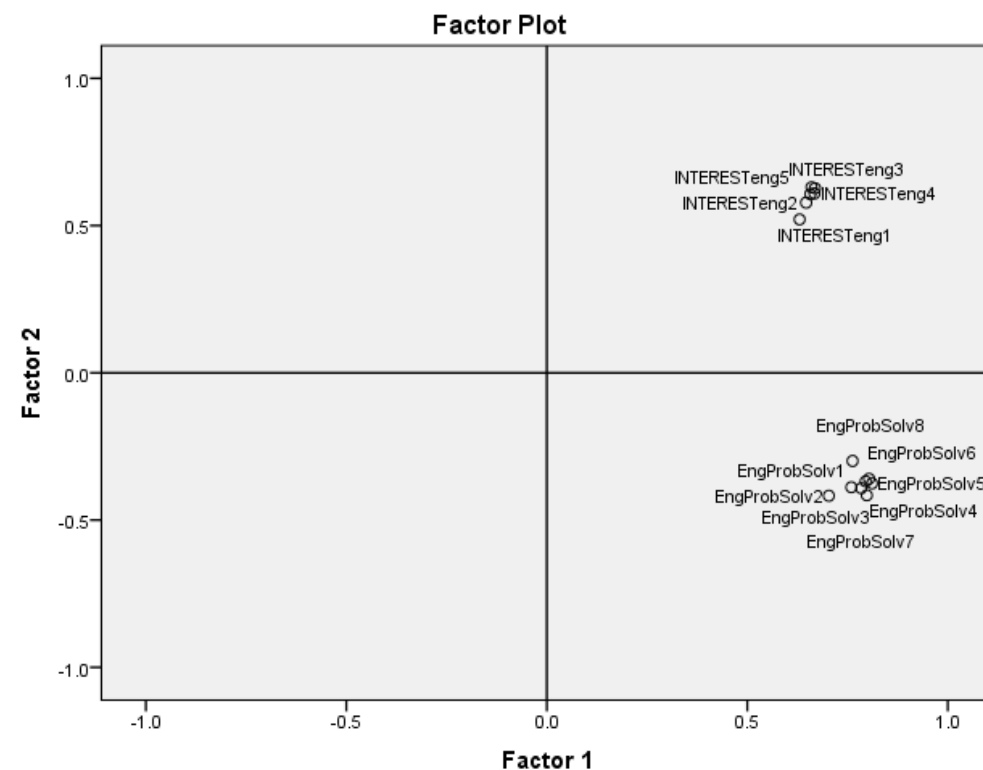
Figure 2.11: Unrotated factor loading plot from Example 1, above.

As with extraction, there are many choices of rotation method, depending on what software you are using. Each uses slightly different algorithms or methods to achieve the same broad goal- simplification of the factor structure. Rotation methods fall into two broad categories: orthogonal and oblique (referring to the angle maintained between the X and Y axes). Orthogonal rotations produce factors that are uncorrelated (i.e., maintain a 90° angle between axes); oblique methods allow the factors to correlate (i.e., allow the X and Y axes to assume a different angle than 90°). Traditionally,

---

remain constant.
[17] Alternatively, you could imagine rotating each cluster of items toward the axis. It really works out to be functionally the same.

researchers have been guided to orthogonal rotation because (the argument went) uncorrelated factors are more easily interpretable. There is also an argument in favor of orthogonal rotation as the mathematics are simpler,[18] and that made a significant difference during much of the 20th century when EFA was performed by hand calculations or much more limited computing power. Orthogonal rotations (which include Quartimax, Equimax, and Varimax) are generally the default setting in most statistical computing packages. Oblique methods include Direct Oblimin and Promax.

**Varimax** rotation seeks to maximize the variance within a factor (within a column of factor loadings) such that larger loadings are increased and smaller are minimized.

Quartimax tends to focus on rows, maximizing the differences between loadings across factors for a particular variable—increasing high loadings and minimizing small loadings.

Equimax is considered a compromise between Varimax and Quartimax, in that it seeks to clarify loadings in both directions.

Promax is recommended by Thompson (2004) as the more desirable oblique rotation choice.[19] It gets its name as a combination of an initial Varimax rotation to clarify the pattern of loadings, and then a procrustean rotation (which is a less common and not discussed here).

Direct Oblimin rotation is another oblique rotation that can sometimes be problematic but often gives very similar results to Promax. Both Promax and Oblimin have parameters that allow the researcher to limit how correlated factors can be (but researchers cannot force factors to be correlated if they are not—in other words, you can limit how strongly correlated the factors are but not the minimum correlation).

The mathematical algorithms for each rotation are different, and beyond the scope of this brief technical note. Note that for all rotations, the goal is the same: simplicity and clarity of factor loadings. For details on how they achieve these goals, you should refer to the manual for your statistical software (e.g., IBM SPSS base statistics manual p. 97,[20] or Gorsuch, 1983; for a good overview of the technical details of different versions of Varimax rotation, see (Forina, Armanino, Lanteri, & Leardi, 1989).

Conventional wisdom in the literature and many texts advises researchers to use orthogonal rotation because it produces more easily interpretable results, but this might a flawed argument. In the social sciences (and many other sciences) we generally expect

---

[18] Researchers also tend to mis-interpret the meaning of "orthogonal" to mean that factor scores are also uncorrelated. Orthogonal factors can (and often do) produce factor scores that are correlated (Nunnally & Bernstein, 1994; Thompson, 2004). More on factor scores in Chapter 7.
[19] However, other authors have argued that there are few substantive differences between the two oblique rotations (Fabrigar et al., 1999).
[20] Retrieved from
ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/22.0/en/client/Manuals/IBM_SPSS_Statistics_Base.pdf

some correlation among factors, particularly scales that reside within the same instrument/questionnaire, regardless of the intentions of the researcher to produce uncorrelated scales (i.e., shared method variance will generally produce nonzero correlations). Therefore using orthogonal rotation results in a loss of valuable information if the factors are really correlated, and oblique rotation should theoretically render a more accurate, and perhaps more reproducible solution.[21] Further, in the unlikely event that researchers manage to produce truly uncorrelated factors, orthogonal and oblique rotation produce nearly identical results, leaving oblique rotation a very low-risk, potentially high benefit choice.

The issue of ease of interpretation is present in one aspect of EFA: when using orthogonal rotation, researchers have only one matrix to interpret. When using oblique rotations, there are two matrices of results to review (described in the next section). In my experience—and in many of our examples—the two matrices tend to parallel each other in interpretation, so again in my mind this does not create an insurmountable barrier.

Factor matrix vs. pattern matrix vs. structure matrix

All extracted factors are initially orthogonal (Thompson, 2004), but remain so only as long as the rotation is orthogonal (we discussed this briefly in chapter 1 regarding PCA). However, even when the factors themselves are orthogonal, factor scores (scores derived from the factor structure; see Chapter 7) are often not uncorrelated despite the factors being orthogonal.

Oblique rotation output is only slightly more complex than orthogonal rotation output. In SPSS output the rotated factor matrix is interpreted after orthogonal rotation; when using oblique rotation we receive both a pattern matrix and structure matrix. This seems to be a source of much confusion in practice.[22] Let's start with a few gentle conceptual definitions:

Factor matrix coefficients are generally reported by most statistical computing packages (like SPSS) regardless of rotation. They represent the unrotated factor loadings, and are generally not of interest.[23]

Pattern matrix coefficients are essentially a series of standardized regression coefficients (betas or **β**s in the regression world) expressing the variable as a function of factor loadings. You can also think of these as the list of ingredients in the recipe (e.g., to make Item 13, add 0.70 of factor 1, 0.13 of factor 2, -0.02 of factor 3, etc. Mix well… delicious!). Like regression coefficients, they hold all other variables (factors) in the equation constant when estimating the pattern matrix coefficients. So, if factors are uncorrelated, pattern and structure coefficients are identical. As factors become more

---

[21] However, some authors have argued that oblique rotations produce *less* replicable results as they might overfit the data to a greater extent. I do not think there is empirical evidence to support this argument, but overfitting the data is a concern to all EFA analyses, as we will discuss further on in the book.

[22] In this section I draw heavily on Thompson (2004), which is always a good reference.

[23] Except for nerds like me trying to understand all this stuff.

---

strongly correlated, the two types of coefficients will become less alike. Thus, think of pattern matrix coefficients as "row" statistics, describing the individual item's unique relationships to each factor.

Structure matrix coefficients are simple correlations between an individual variable and the composite or latent variable (factor). In multiple regression, these would be similar to correlations between an individual variable and the predicted score derived from the regression equation. The difference between structure and pattern coefficients are the difference (again, returning to regression) between simple correlations and semipartial (unique relationship only) correlations.

Pattern vs. Structure matrix. If all factors are perfectly uncorrelated with each other, the pattern and structure coefficients are identical. When factors are uncorrelated there is no effect of holding other factors constant when computing the pattern matrix, and the structure and pattern coefficients would be the same, just like simple and semipartial correlations in multiple regression with perfectly uncorrelated predictors.

Thompson (2004; see also Gorsuch, 1983; Nunnally & Bernstein, 1994) and others have argued that it is essential to interpret both pattern and structure coefficients in order to correctly and fully interpret the results of an EFA. In practice, few do. Further, when rotations are oblique and factors are correlated, they argue it is important to report the intercorrelations of factors also. I will highlight this process when appropriate.

Rotation example 1: Engineering data

I have reproduced the original unrotated factor loadings for your convenience in Table 2.5. As you can see in Table 2.5, although we expected (and see evidence of) two very clear factors, the factor loadings are not immediately identifiable as two separate factors prior to rotation. To be sure, looking at the unrotated Factor 1 loadings, all fourteen items seem to be similar. It is only in combination with the loadings on Factor 2 where the two factors separate. If one plots each item in two-dimensional space (Factor 1 on the X axis, and Factor 2 on the Y axis), we see clearly the separation, as presented a couple pages back in Figure 2.11.

As you examine Figure 2.11 and imagined rotating the axes so that they intersected more closely with the center of each cluster of variables, I hope you can visualize turning the axes so that they come into alignment. You might be wondering what that does for us in terms of clarifying the factor loadings. Once we make that rotation, the factor pattern coefficients also have now changed, with some getting larger, and some getting smaller. As you can see in Figure 2.12 (and Table 2.5), following rotation of the axes (or items), the items now fall closely about each axis line. This has the effect of making the factor loading pattern much clearer as one of the two pairs of coordinates for each item tends to be close to 0.00, as you can see in Table 2.5. In this example analysis, the factors were correlated $r = 0.37$.

Table 2.5

Factor loading matrix from example #1 (engineering data)

Variable:	Unrotated		Rotated Pattern Matrix		Rotated Structure Matrix	
	1	**2**	**1**	**2**	**1**	**2**
EngProbSolv1	.759	-.389	.859	-.016	.853	.300
EngProbSolv2	.703	-.418	.841	-.071	.815	.239
EngProbSolv3	.784	-.392	.879	-.008	.877	.316
EngProbSolv4	.798	-.416	.909	-.025	.900	.310
EngProbSolv5	.811	-.375	.886	.021	.894	.347
EngProbSolv6	.795	-.369	.869	.020	.876	.340
EngProbSolv7	.804	-.360	.868	.033	.880	.352
EngProbSolv8	.763	-.299	.790	.072	.816	.362
INTERESTeng1	.630	.521	.042	.801	.337	.817
INTERESTeng2	.660	.630	-.023	.921	.316	.912
INTERESTeng3	.669	.627	-.014	.922	.325	.917
INTERESTeng4	.668	.609	-.001	.904	.332	.904
INTERESTeng5	.657	.607	-.007	.897	.324	.894
INTERESTeng6	.647	.578	.009	.864	.327	.867

Note: Principal Axis Factoring extraction, Oblimin rotation. Correlation between two factors is 0.37.
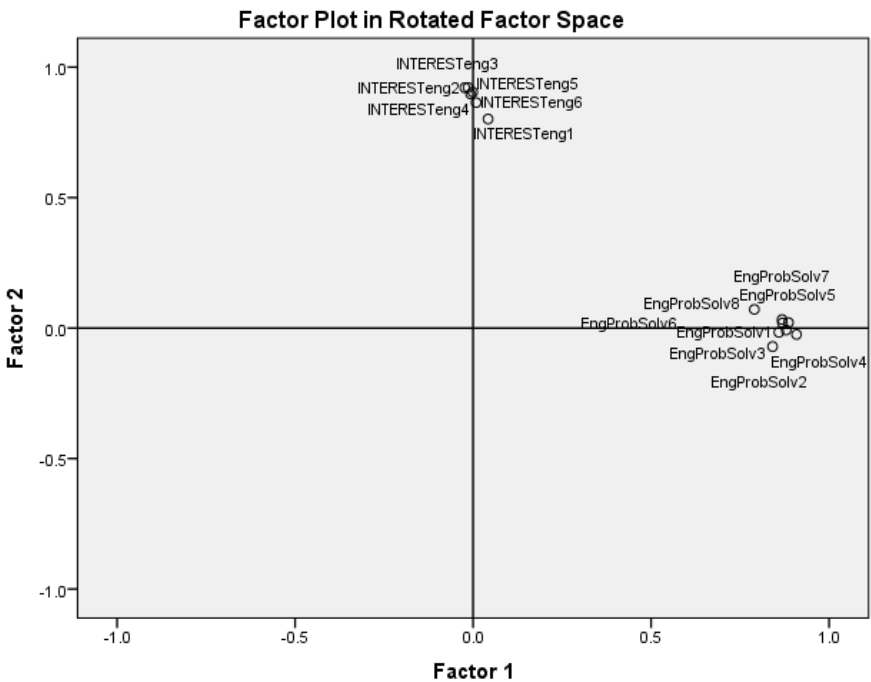
Figure 2.12: Rotated factor solution following Principal Axis Factoring extraction and oblique (Oblimin) rotation

Most statistical packages will allow small loadings to be suppressed following rotation, so that the results become even more obvious and immediately apparent. I chose to keep all loadings in the table but to highlight the rotated loadings that were of interest.

Do orthogonal and oblique rotations produce noticeable differences? Orthogonal and oblique rotations will produce virtually identical solutions in the case where factors are perfectly uncorrelated. As the correlation between latent variables diverges from $r = 0.00$, then the oblique solution will produce increasingly clearer results. Looking at the same data after orthogonal (Varimax) rotation presented in Figure 2.13, one can see that this outcome still provides a similar but less ideal solution. This is because these factors are modestly correlated, but the mandate to maintain a 90º angle between axes means that the centroids of the clusters cannot move closer to the axis lines. In this case, the difference is not great, but noticeable. This is a small but clear example of the higher efficacy of oblique rotations to create clear patterns of results in EFA where the factors are indeed correlated. Given that there is no rationale I am aware of for using orthogonal rotation instead of oblique rotation (except tradition), there is no reason in my mind to accept sub-optimal rotation by insisting on using orthogonal rotations.
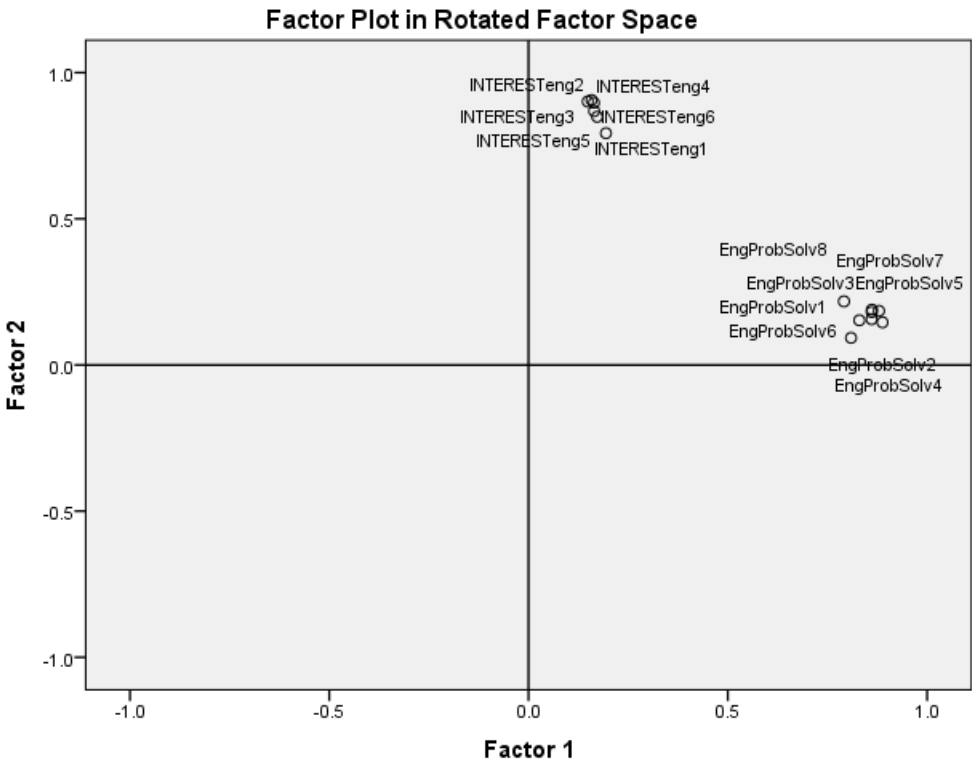
Figure 2.13: Engineering data using orthogonal rotation (Varimax)

Rotation example 2: Self-Description Questionnaire data

To explore this example, I performed ML extraction of three factors (in accord with several criteria reviewed above) with Promax rotation. Interestingly, these scales were minimally correlated ($r = 0.15, 0.22, 0.26$), but the solution is what we would expect, given the theoretical model the scale was developed with. As you can see in Table 2.6, the final rotated solution gives us a clear, theoretically consistent factor structure with subscale items aligning as expected.

Table 2.6
Unrotated and rotated factor loadings for SDQ data

Var:	Unrotated Factor Loadings			Rotated Pattern Matrix			Rotated Structure Matrix		
	1	2	3	1	2	3	1	2	3
Math1	.802	.088	-.373	.901	-.040	-.037	.887	.152	.088
Math2	.810	.070	-.298	.863	.008	.012	.866	.203	.144
Math3	.828	.083	-.301	.881	-.001	.023	.885	.201	.155
Math4	-.572	-.004	.212	-.601	-.049	.021	-.608	-.177	-.081
Par1	.360	-.524	.349	.002	.718	.023	.166	.725	.209
Par2	-.252	.533	-.293	.060	-.680	.052	-.084	-.654	-.114
Par3	.426	-.613	.370	.028	.827	-.002	.212	.833	.216
Par4	-.359	.388	-.335	-.036	-.583	-.100	-.181	-.617	-.256
Par5	.367	-.568	.315	.018	.749	-.030	.181	.746	.166
Eng1	.348	.309	.634	-.005	.031	.779	.119	.231	.786
Eng2	.310	.419	.636	-.016	-.082	.842	.092	.131	.818
Eng3	.406	.378	.644	.052	-.017	.845	.175	.212	.848
Eng4	-.257	-.179	-.552	.060	-.102	-.609	-.054	-.245	-.626

Note: factors correlations ranged from 0.15 to 0.26.

In summary, this second example is about as good as one can hope for when performing an exploratory factor analysis: the theoretical structure matches the majority of the recommendations from the factor extraction criteria, and in the end, the rotated solution aligns with what was hoped for. Indeed, even the communalities are reasonably strong and the analysis explained 59.37% of the overall variance, which is relatively good. Note that the pattern matrix and structure matrix are similar because the correlations between the factors are minimal.

Rotation example 3: Geriatric Depression Scale data

Given that the data from the GDS was less clear in terms of factor structure, we are going to have to experiment and explore more than previous examples. Let us remember that we want to favor parsimonious solutions. Also, because all proposed subscales relate to depression in some way, they should be correlated. We will therefore use an oblique rotation while exploring.

Below is a one-factor solution (PAF extraction), which would be defensible based

on the single large eigenvalue and MAP criteria. This single factor accounts for 23.80% of the variance and communalities that ranged from 0.11 to 0.47, which is not a strong result. As you can see, many loadings are low, and even the highest loadings are in the .60-.70 range. This is likely due to the poor measurement (0, 1 only).

Table 2.7
Factor loadings for GDS data with one factor and five factors extracted

	One-factor	Five-factor pattern loadings				
	1	1	2	3	4	5
GDS01	.591	.441		.302		
GDS02	.406		.452			
GDS03	.629			.509		
GDS04	.534				.719	
GDS05	.512	.418				
GDS06	.524			.373		
GDS07	.473	.694				
GDS08	.401					
GDS09	.607	.782				
GDS10	.618			.405		
GDS11	.429				.541	
GDS12	.356		.452			
GDS13	.418				.301	
GDS14	.231					.563
GDS15	.467	.647				
GDS16	.684			.399		
GDS17	.661		.361	.454		
GDS18	.421			.355		
GDS19	.566		.499			
GDS20	.429		.496			
GDS21	.448		.599			
GDS22	.552			.756		
GDS23	.418			.452		
GDS24	.400				.508	
GDS25	.578			.434		
GDS26	.460					.492
GDS27	.376	.343				
GDS28	.432		.435			
GDS29	.327					
GDS30	.331					.470

Note: Factor loadings less than 0.30 suppressed in the five-factor model for ease of interpretation. Recall that one-factor solutions are not rotated so the loadings might seem sub-optimal.

When the theoretically-supported five factors are extracted (Table 2.7), 37.02% of

the variance is accounted for, and the communities range from 0.15 to 0.64- better but not impressive. Further, the items loading on various factors do not match the theoretical framework, and thus do not (to me) make sense. If it does not make sense in an alternative way, I would be reluctant to put it forward as the best model.

Table 2.8
Eight factor solution for GDS data.

Variable:	Factor pattern loadings							
	1	2	3	4	5	6	7	8
GDS01								.653
GDS02			.497					.315
GDS03		.352						.412
GDS04				.767				
GDS05	.372							
GDS06					.429			
GDS07	.740							
GDS08					.527			
GDS09	.796							
GDS10		.379						
GDS11				.606				
GDS12							.877	
GDS13					.665			
GDS14						.679		
GDS15	.605							
GDS16		.402						
GDS17		.501	.378					
GDS18		.318			.372			
GDS19			.408					
GDS20			.541					
GDS21			.565					
GDS22		.735						
GDS23		.456						
GDS24				.334				
GDS25		.593						
GDS26						.494		
GDS27			.312					
GDS28							.438	
GDS29			.366					
GDS30			.362			.401		

Note: pattern coefficients reported to conserve space.

As you can see in Table 2.8, the next possibility is an eight-factor solution, which was indicated not only by the Kaiser criterion (eigenvalues greater than 1.0) but also by parallel analysis. With eight factors extracted, the model accounted for 43.44% of the

variance, and communalities ranged from 0.17 to 0.67, slightly better than the five factor model (but to my mind not any more interpretable). For the sake of space I will leave the three-factor model to you to examine. It might be conceptually intriguing.

Summary of rotation examples. In this section we examined the concept of rotation- the purpose of rotation, what actually rotates, and several different methods for rotation. Finally, we examined the results of the analyses after rotation. When the data were clear with a strong factor structure (as in example #1 and 2, with the engineering and SDQ data), almost any rotation will do a good job clarifying the factor structure, but I argued that oblique rotations did slightly better. Further, the pattern and structure coefficients were reported, as this was an oblique rotation, but the results were clear in both cases, and thus, the oblique rotation did not overly complicate the results.

In the third example (GDS data), the way forward was less clear. We explored a single-factor model, which may ultimately be the most desirable given our preference for parsimony. However, in this model, the communalities were lower than ideal, and the overall variance accounted for was relatively low. Guidelines we previously explored recommended extracting either 3, 5, or 8 factors, but none of them seemed to make more sense to me over the single factor model (you will have to decide if the three-factor solution is the best or not). This scale might need a larger sample, revision, or to be examined in the context of confirmatory methods in order to determine which might model be superior. However, we cannot just be guided by empirical data. The latent variables we construct have to make sense.

Going back to the eight-factor model, the fifth factor (items 6, 8, 13, 18) focuses on intrusive thoughts, anxiety, and worry. This is an example of a factor that could make conceptual sense and be defensible. [24] Likewise, the first factor (items 5, 7, 9, and 15) all have hopeful, positive concepts and as such also make sense. So one could label this some aspect of positivity or hopefulness, except that other items, like 19, 21, and 27 would also fall into that category yet load on the third factor. In the five-factor model this positivity factor comes together more, but does not get there all the way.

At the end of the day, EFA is about empirically constructing a narrative that makes sense. See if you can come up with one of the models that makes more sense than simply saying all the items measure depression.

Standard practice in social science

One of the challenges to utilizing best practices is that for three of the important EFA analytic decisions (extraction, rotation, and number of factors to retain), the defaults in common statistical packages (e.g., SPSS, SAS) are not ideal. These defaults often include (a) principal components as the extraction method of choice,[25] (b) extract the number of factors according to the Kaiser (1970) criterion (retaining factors with eigenvalues greater than 1.0), and (c) use Varimax for rotation of factors (which requires factors to be orthogonal, or completely uncorrelated with each other.

---

[24] I am not a clinically trained psychologist so I am looking from a naïve perspective- apologies to the experts out there if I miss some nuance.
[25] I won't beat this PCA vs. EFA drum again.

Unfortunately, software defaults tend to drive analysis decision-making, which is probably why PCA extraction, the "Little Jiffy" criterion for number of factors to keep, and Varimax rotation are the most common decisions made in articles published in many sciences. This has been confirmed by many authors in many fields. Two colleagues and I (Osborne, Costello, & Kellow, 2008) published a review of articles using EFA in two prominent measurement journals: Educational and Psychological Measurement (EPM) and Personality and Individual Differences (PID) over a six-year period. These journals were chosen because of their prominence[26] in the field of measurement and the prolific presence of EFA articles within their pages. After screening out studies that employed only confirmatory factor analysis or examined the statistical properties of EFA or CFA approaches using simulated data sets, a total of 184 articles were identified, reporting on 212 distinct EFA analyses. Information extracted from the EFA articles were: a) factor extraction methods; b) factor retention rules; c) factor rotation strategies; and d) saliency criteria for including variables.

Factor Extraction Methods. Almost two-thirds of all researchers (64%) in these journals used principal components (PCA). The next most popular choice was principal axis factoring (PAF) (27%). Techniques such as maximum likelihood were infrequently invoked (6%). A modest percentage of authors (8%) utilized multiple methods on their data to compare the results for similar structure.

Factor Extraction Rules. The most popular method used for deciding the number of factors to retain was the Kaiser (1960, 1970) criterion of eigenvalues greater than 1.0 (45%). An almost equal proportion used the scree test (42%). Use of other methods, such as parallel analysis, was comparatively infrequent (about 8%). Many authors (41%) explored multiple criteria for factor retention. Among these authors, the most popular choice was a combination of the Kaiser Criterion and scree methods (67%).

Factor Rotation Strategies. As expected, Varimax rotation was most often employed (47%), with Oblimin being the next most common (38%). Promax (another oblique rotation) also was used in 11% of analyses. A number of authors (18%) employed both Varimax and Oblimin solutions to examine the influence of correlated factors on the resulting factor pattern/structure matrices.

Saliency Criteria for Including Variables. Thirty-one percent of EFA authors did not articulate a specific criterion for interpreting salient pattern/structure coefficients, preferring instead to examine the matrix in a logical fashion, considering not only the size of the pattern/structure coefficient, but also the discrepancy between coefficients for the same variable across different factors (components) and the logical "fit" of the variable with a particular factor.

Of the 69% of authors who identified an a priori criterion as an absolute cutoff, 27% opted to interpret coefficients with as absolute value of 0.30 or higher, while 24% chose a 0.40 cutoff. Other criteria chosen with modest frequency (both about 6%) included 0.35 and 0.50 as absolute cutoff values with the rest ranging from the marginally defensible 0.25 to the almost indefensible 0.80.

---

[26] Which, in our reasoning, should lead to the most progressive and rigorous methodology

Summary. Not surprisingly, the hegemony of default settings in major statistical packages continues to dominate the pages of EPM and PID. The "Little Jiffy" model espoused by Kaiser (1960, 1970), which combined PCA with Varimax rotation and retention of all factors with an eigenvalue greater than 1.0, is alive and well. It should be noted that this situation is almost certainly not unique to EPM or PID authors. Another survey of a recent two-year period in PsycINFO (reported in Costello & Osborne, 2005) yielded over 1700 studies that showed similar results. Informal perusal of the top journals from other empirical fields easily confirms the prevalence of this situation as current practice.

There are probably good historical reasons why these defaults and practices have become ingrained in the factor analysis culture. When they were promulgated in the middle and later parts of the 20th century, they were solid methods with no better alternatives, given the state of statistical computing. The "Little Jiffy" method will often yield acceptable results that will generalize, but a significant amount of subsequent research points to the fallibility of this methodology. The goal of this book is to help you apply the best evidence-based practices available today, in the 21st century. Of course, we must constantly keep in mind that this is an exploratory technique, and as such, should be interpreted accordingly. No EFA or PCA should ever be considered the last word in examination of a research instrument. Confirmatory methods are designed to much more clearly and rigorously test models we propose.

Chapter 2 Summary

There are two basic aspects to setting up the factor analysis that we attended to in this chapter: extraction and rotation. Many of us, even those with years of experience using EFA, remain unclear on some of the nuances and details of what exactly is happening "under the hood" when we perform this analysis. Sticking with the default settings in most modern statistical packages will generally *not lead to using best practices.* In SPSS, for example, the default extraction is PCA, and the default rotation is Varimax. Both are solid choices if you are a psychologist in the 1960s, but in the 21st century, we can do better.

Many scholars have written on guidelines for extraction and rotation of factors, focusing on eigenvalues, scree plots, parallel analysis, replication, and so on. It is my belief that the over-arching value has to be theoretical framework and an easily-interpretable factor structure. Absent this, which we use to make sense of data, none of the technical details seem important.

Note that in this chapter I used two-dimensional plots to illustrate the example. If an instrument is uni-dimensional no rotation is possible. If an instrument is three- (or more) dimensional, then items are plotted in multidimensional space, and three (or more) axes are rotated within this multidimensional space with the same goals.

Chapter 2 Exercises

1. Download the engineering, SDQ, and/or GDS data from the book web site on http://jwosborne.com. Replicate the results reported in the chapter. Specifically:
    a. Examine extraction methods to see which one provides the best extraction.
    b. Explore application of MAP and parallel analysis for each of the data sets (SPSS and SAS syntax available on web site).
    c. Examine factor loading plot prior to rotation (if you examine the factor loading plot for the GDS it will be a three-dimensional plot that is more complex to interpret) and after trying various rotation strategies. Which do you think provides the best clarification of the result?
    d. For the GDS data, examine a 3-factor solution to see if it is more sensible than the one-, five- or eight-factor solutions described in the chapter. Do the factors make sense? If so, describe what each latent variable is measuring.
2. Download and perform EFA on data from an early version of one of my early studies on identification with academics (Osborne, 1997).[27] This was intended to be a measure of identification with academics (the extent to which a student defines oneself as a student as part of self-concept). There were supposed to be three scales, measured on a scale of 1 (disagree strongly) to 7 (agree strongly):
    a. centrality of academics to self (items 1, 4, 7, 9, 10, 11, 15, 16, 18, 19, 20, 21)
    b. feelings of discrimination (2, 3, 6)
    c. diagnosticity of academic outcomes (5, 8, 12, 13, 14, 17)

---

[27] Note that this is not the final version of the scale that was used in the publication, which was a single-factor scale. This is a pilot version of the three-factor version that was not published.

3 SAMPLE SIZE MATTERS

Larger samples are better than smaller samples (all other things being equal) because larger samples tend to minimize the probability of errors, maximize the accuracy of population estimates, and increase the generalizability of the results. Unfortunately, there are few sample size guidelines for researchers using EFA or PCA, and many of these have minimal empirical evidence (e.g., Guadagnoli & Velicer, 1988).

This is problematic because statistical procedures that create optimized linear combinations of variables (e.g., multiple regression, canonical correlation, and EFA) tend to "overfit" the data. This means that these procedures optimize the fit of the model to the given data; yet no sample is perfectly reflective of the population. Thus, this overfitting can result in erroneous conclusions if models fit to one data set are applied to others. In multiple regression this manifests itself as inflated R^2 (shrinkage) and mis-estimated variable regression coefficients (Cohen, Cohen, West, & Aiken, 2002, pp. 83-84). In EFA this "overfitting" can result in erroneous conclusions in several ways, including the extraction of erroneous factors or mis-assignment of items to factors (e.g., Tabachnick & Fidell, 2001, pp., p. 588).

Published sample size guidelines.

In multiple regression texts some authors (e.g., Pedhazur, 1997, p. 207) suggest subject to variable ratios of 15:1 or 30:1 when generalization is critical. But there are few explicit guidelines such as this for EFA (e.g., Baggaley, 1983). Two different approaches have been taken: suggesting a minimum total sample size, or examining the ratio of parameters such as subjects to variables, as in multiple regression.

Comfrey and Lee (1992) suggest that "the adequacy of sample size might be evaluated very roughly on the following scale: 50 – very poor; 100 – poor; 200 – fair; 300 – good; 500 – very good; 1000 or more – excellent" (p. 217). Guadagnoli and Velicer (1988) review several studies that conclude that absolute minimum sample sizes, rather than subject to item ratios, are more relevant. These studies range in their recommendations from an N of 50 (Barrett & Kline, 1981) to 400 (Aleamoni, 1976). In my mind some of these recommendations are ridiculous, as they could result in analyses estimating far more parameters than available subjects.

The case for ratios. There are few scholars writing about multiple regression camp who would argue that total N is a superior guideline than the ratio of subjects to variables, yet authors focusing on EFA occasionally vehemently defend this position. It is interesting precisely because the general goal for both analyses is similar: to take individual variables and create optimally weighted linear composites that will generalize to other samples or to the population. While the mathematics and procedures differ in the details, the essence and the pitfalls are the same. Both EFA and multiple regression risk over-fitting of the estimates to the data (Bobko & Schemmer, 1984), both suffer from lack of generalizability particularly keenly when sample size is too small.

Absolute sample sizes seem simplistic given the range of complexity factor analyses can exhibit-- each scale differs in the number of factors or components, the number of items on each factor, the magnitude of the item-factor correlations, and the correlation between factors, for example. This has led some authors to focus on the ratio of subjects to items, or more recently, the ratio of subjects to parameters (as each item will have a loading for each factor or component extracted), as authors do with regression, rather than absolute sample size when discussing guidelines concerning EFA.

Gorsuch (1983, p.332) and Hatcher (1994, p. 73) recommend a *minimum* subject to item ratio of at least 5:1 in EFA, but they also describe stringent guidelines for when this ratio is acceptable, and they both note that higher ratios are generally better. There is a widely-cited rule of thumb from Nunnally (1978, p. 421) that the subject to item ratio for exploratory factor analysis should be at least 10:1, but that recommendation was not supported by empirical research. Authors such as Stevens (2002) have provided recommendations ranging from 5-20 participants per scale item, with Jöreskog and Sörbom (1996) encouraging at least *10 participants per parameter estimated*.

There is no one ratio that will work in all cases; the number of items per factor and communalities and item loading magnitudes can make any particular ratio overkill or hopelessly insufficient (MacCallum, Widaman, Preacher, & Hong, 2001).

Are subject: item ratios an important predictor of good EFA analyses?

Unfortunately, much of the literature that has attempted to address this issue, particularly the studies attempting to dismiss subject to item ratios, use flawed data. I will purposely not cite studies here to protect the guilty, but consider it sufficient to say that many of these studies either tend to use highly restricted ranges of subject to item ratios or fail to adequately control for or vary other confounding variables (e.g., factor loadings, number of items per scale or per factor/component) or restricted range of N. Some of these studies purporting to address subject to item ratio *fail to actually test subject to item ratios* in their analyses.

Researchers seeking guidance concerning sufficient sample size in EFA are left between two entrenched camps-- those arguing for looking at total sample size and those looking at ratios.[28] This is unfortunate, because both probably matter in some sense, and ignoring either one can have the same result: errors of inference. Failure to have a representative sample of sufficient size results in unstable loadings (Cliff, 1970), random, non-replicable factors (Aleamoni, 1976; Humphreys, Ilgen, McGrath, &

---

[28] And of course, those who don't consider sample size at all when planning their research.

Montanelli, 1969), and lack of generalizability to the population (MacCallum, Widaman, Zhang, & Hong, 1999).

Sample size in practice. If one were to take either set of guidelines (e.g, 10:1 ratio or a minimum N of 400 - 500) as reasonable guidelines, a casual perusal of the published literature shows that a large portion of published studies come up short. One can easily find articles reporting results from EFA or PCA based on samples with fewer subjects than items or parameters estimated that nevertheless draw substantive conclusions based on these questionable analyses. Many more have hopelessly insufficient samples by either guideline.

One survey by Ford, MacCallum, and Tait (1986) examined common practice in factor analysis in industrial and organizational psychology during the ten year period of 1974 - 1984. They found that out of 152 studies utilizing EFA or PCA, 27.3% had a subject to item ratio of less than 5:1 and 56% had a ratio of less than 10:1. This matches the perception that readers of social science journals get, which is that often samples are too small for the analyses to be stable or generalizable.

I and my colleagues published the results of a survey of current practices in the social sciences literature (Osborne et al., 2008). In this survey, we sampled from two years' (2002, 2003) worth of articles archived in PsycINFO that both reported some form of EFA and listed both the number of subjects and the number of items analyzed (303 total articles surveyed). We decided the best method for standardizing our sample size data was via subject to item ratio, since we needed a criterion for a reasonably direct comparison to our own data analysis. The results of this survey and are summarized in Table 3.1. A large percentage of researchers report factor analyses using relatively small samples. In a majority of the studies (62.9%) researchers performed analyses with subject to item ratios of 10:1 or less. A surprisingly high proportion (almost one-sixth) reported factor analyses based on subject to item ratios of only 2:1 or less (note that in this case there would be more parameters estimated than subjects if more than 1 factor is extracted).

Table 3.1:
Current practice in factor analysis in 2002-03 Psychology journals

Subject to item ratio	% of studies	Cumulative %
2:1 or less	14.7%	14.7%
> 2:1, ≤ 5:1	25.8%	40.5%
> 5:1, ≤ 10:1	22.7%	63.2%
> 10:1, ≤ 20:1	15.4%	78.6%
> 20:1, ≤100:1	18.4%	97.0%
> 100:1	3.0%	100.0%

A more recent survey of EFA practices in *Educational and Psychological Measurement*, *Journal of Educational Psychology*, *Personality and Individual Differences*, and *Psychological Assessment* by Henson and Roberts' (2006) indicates a median sample size of 267 for

reported EFAs, mean subject : item ratio of 11, and a median of 60 parameters (20 items x 3 factors) estimated. As you will see below, these are not comforting statistics. Given the stakes and the empirical evidence on the consequences of insufficient sample size, this is not exactly a desirable state of affairs.

Size matters two different ways

This section focuses on one particularly interesting and relatively well-executed study on this issue—that of Guadagnoli and Velicer (1988). In this study, the authors used Monte Carlo methods to examine the effects of number of factors (3, 6, 9, 18), the number of variables (36, 72, 108, and 144), average item loadings (.40, .60, or .80), and number of subjects (*Ns* of 50, 100, 150, 200, 300, 500, and 1000) on the stability of factor patterns in EFA. In these data each item loaded on only one factor, all items loaded equally on every factor, and each factor contained an equal number of variables. Their study represents one of the few studies to manipulate all of these important aspects across a reasonable range of variation seen in the literature (with the two possible exceptions: first, people often have less than 36 items in a scale, and second, the factor loading patterns are rarely as clear and homogenous as in these data).

Guadagnoli and Velicer's (1988) study was also interesting in that they used several different high-quality fit/agreement indices. Equally interesting is the authors' strong assertion that total sample size is critical, although they never actually operationalize subject to item ratio, nor test whether total *N* is a better predictor of important outcomes than subject to item ratio, although given their data it was possible to do so. Thus, Costello and I (Osborne & Costello, 2004) re-analyzed their published data to examine whether total sample size, or sample size per parameter estimated (as is more reasonable) produced the most important indicators of quality analyses.

As previous research has reported, strong factor loadings led to better indicators (e.g., less discrepancy between population and sample results, and the odds of getting the correct component pattern increased dramatically). Unfortunately, the magnitude of item loadings is not realistically within the control of the researcher.

Contrary to prior studies, neither the absolute number of variables nor total sample size (*N*) had a significant unique effect when all other aspects of the analysis were considered. Total N was significant in all analyses until ratios were taken into account, at which point they became non-significant. The ratio of subjects to items had a significant and substantial influence on several outcomes, such as improved match between sample and population results, and the odds of getting a correct factor pattern matrix increased.

Costello and Osborne (2005) analyses

While the data from Guadagnoli and Velicer (1988) are illuminating, one frustration is the unrealistically clean nature of the data. Real data are messier than that, and we wanted to replicate and extend the findings from these artificial data with real data. Costello and I (2005) used data similar to that used for Example 2 in the previous chapter-- students who completed Marsh's Self-Description Questionnaire (SDQ II; Marsh, 1990) in the NELS 88 data set (Curtin, Ingels, Wu, & Heuer, 2002).[29]

To explore the effects of sample size, we drew samples (with replacement between samplings), extracting twenty samples of sizes ranging from 2:1, 5:1, 10:1, and 20:1 subject to item ratios (creating sample sizes of N = 26, 65, 130, and 260 respectively). The samples drawn from the population data were analyzed using maximum likelihood extraction with Direct Oblimin rotation. For each sample, the magnitude of the eigenvalues, the number of eigenvalues greater than 1.0, the factor loadings of the individual items, and the number of items incorrectly loading on a factor were recorded. In order to assess accuracy as a function of sample size, we computed average error in eigenvalues and average error in factor loadings. We also recorded aberrations such as occasions when a loading exceeds 1.0, and instances of failure for ML to converge on a solution after 250 iterations.

Finally, a global assessment of the correctness or incorrectness of the factor structure was made. If a factor analysis for a particular sample produced three factors, and the items loaded on the correct factors (the same structure we explored in the previous chapter), that analysis was considered to have produced the correct factor structure (i.e., a researcher drawing that sample, and performing that analysis, would draw the correct conclusions regarding the underlying factor structure for those items). If a factor analysis produced an incorrect number of factors with eigenvalues greater than 1.0 (some produced up to 5), or if one or more items failed to load on the appropriate factor, that analysis was considered to have produced an incorrect factor structure (i.e., a researcher drawing that sample, and performing that analysis, would not draw the correct conclusions regarding the underlying factor structure).

Sample size. In order to examine how sample size affected the likelihood of errors of inference regarding factor structure of this scale, an analysis of variance was performed, examining the number of samples producing correct factor structures as a function of the sample size. The results of this analysis are presented in Table 3.2. As expected, larger samples tended to produce solutions that were more accurate. Only 10% of samples in the smallest (2:1) sample produced correct solutions (identical to the population parameters), while 70% in the largest (20:1) produced correct solutions. Further, the number of misclassified items was also significantly affected by sample size. Almost two of thirteen items on average were misclassified on the wrong factor in the smallest samples, whereas just over one item in every two analyses were misclassified in the largest samples. Finally, two indicators of trouble—the presence of factor loadings greater than 1.0, and/or failure to converge, were both exclusively observed in the smaller samples, with almost one-third of analyses in the smallest sample size category failing to produce a solution.

What is particularly illuminating is to go back to Table 3.1, noting that while the majority of recent papers have subject: item ratios in the lower ranges, in our analyses the error rates for these ranges are extraordinarily high. Specifically, approximately two-thirds of published EFA studies have subject: item ratios of less than 10:1, while at the same time this ratio is associated with an error rate of approximately 40%.

Table 3.2

[29] NELS 88 data and information is available from the IES web site: http://nces.ed.gov/surveys/nels88/

The effects of subject to item ratio on exploratory factor analysis

Variable:	2:1	5:1	10:1	20:1	$F_{(3,76)}$
% samples with correct structure	10%	40%	60%	70%	13.64*** (.21)
Average number of items misclassified on wrong factor	1.93	1.20	0.70	0.60	9.25*** (.16)
Average error in eigenvalues	.41	.33	.20	.16	25.36*** (.33)
Average error in factor loadings	.15	.12	.09	.07	36.38*** (.43)
% fail to converge after 250 iterations	30%	0%	0%	0%	8.14*** (.24)
% with loadings >1	15%	20%	0%	0%	2.81* (.10)

*Note: η2 reported in parentheses for significant effects, * p < .05, *** p < .0001*

Chapter 3 Summary: Does sample size matter in EFA?

The goal of this chapter was to summarize some of the scholarship surrounding the age-old question of "how large a sample is large enough?" Recall from Chapter 2 that the SDQ had a very strong and clear factor structure, at least in a large sample. Unfortunately, these results suggest that EFA is an error-prone procedure even when the scale being analyzed has a strong factor structure, and even with large samples. Our analyses demonstrate that at a 20:1 subject to item ratio there are error rates well above the field standard alpha = .05 level.

This again reinforces the point that EFA is *exploratory*. It should be used *only* for exploring data, not hypothesis or theory testing, nor is it suited to "validation" of instruments. I have seen many cases where researchers used EFA when they should have used confirmatory factor analysis. Once an instrument has been developed using EFA and other techniques, it is time to move to confirmatory factor analysis to answer questions such as "does an instrument have the same structure across certain population subgroups?" Based on the data presented in this chapter, I think it is safe to conclude that researchers using large samples and making informed choices from the options available for data analysis are the ones most likely to accomplish their goal: to come to conclusions that will generalize beyond a particular sample to either another sample or to the population (or *a* population) of interest. To do less is to arrive at conclusions that are unlikely to be of any use or interest beyond that sample and that analysis.

Chapter 3 Exercises

1. Experiment with our Marsh SDQ data set (or another large data set you have available). Using the results of the EFA from the entire sample, draw small random samples of varying subject: item ratios representing:
   a. 2:1
   b. 5:1
   c. 10:1
   d. 20:1

   Explore the effect of having an inappropriately small sample on the goodness of the solution. Do the results of the small samples replicate the results from the large sample "population"?

2. Review EFA analyses in top journals in your field. What subject: item ratios do you find in these articles? Are they sufficient, given the results from this chapter and your experiments in #1?

4 REPLICATION STATISTICS IN EFA

"Factor analysis is really not concerned with exactness, only good approximation."
-Nunnally & Bernstein, 1994, p. 509

I have repeatedly recommended that readers and researchers to keep in mind the exploratory nature of EFA- a procedure that by nature is quirky, temperamental, valuable, and interesting. As we discussed in Chapter 3, exploratory factor analysis takes advantage of all the information in the interrelationships between variables, whether those interrelationships are representative of the population of not. In other words, EFA tends over-fit a model to the data such that when the same model is applied to a new sample, the model is rarely as good a fit. When we as readers see a single EFA, often on an inadequate sample (as discussed in Chapter 3), we have no way of knowing whether the results reported are likely to generalize to a new sample or to the population. But it seems as though this might be useful information.

Why replication is important in EFA

If you read enough articles reporting the results from factor analyses, too often you will find confirmatory language used regarding exploratory analyses. We need to re-emphasize in our discipline that EFA is *not* a mode for testing of hypotheses or *confirming* ideas (e.g., Briggs & Cheek, 1986; Floyd & Widaman, 1995), but rather for exploring the nature of scales and item inter-relationships. EFA merely presents a solution based on the available data.

These solutions are notoriously difficult to replicate, even under abnormally ideal circumstances (exceptionally clear factor structure, very large sample to parameter ratios, strong factor loadings, and high communalities). As mentioned already, many point estimates and statistical analyses vary in how well they will generalize to other samples or populations (which is why we are more routinely asking for confidence intervals for point estimates). But EFA seems particularly problematic in this area.

We find this troubling, and you should too. Of course, we have no specific information about how replicable we should expect particular factor structures to be because direct tests of replicability are almost never published. As Thompson (1999) and others note, replication is a key foundational principle in science, but we rarely find replication studies published. It could be because journals refuse to publish them, or because researchers don't perform them. Either way, this is not an ideal situation.

Let's bring replication to EFA.

Authors can (and, I argue, should) directly estimate the replicability of their exploratory factor analyses reported in scientific journals. Authors (e.g., Thompson, 2004) have introduced replicability procedures for EFA, similar to those procedures considered best practices in validation of prediction equations in multiple regression (Osborne, 2000, 2008a). Although few authors perform the procedure, I hope you will see the intuitive appeal.

Specifically, since the goal of EFA is usually to infer or explore the likely factor structure of an instrument when used within a particular population, it is important to know whether a factor structure within a particular data set is likely to be observed within another, similar data set.[30] The lowest threshold for replicability should be replicating the same basic factor structure (same number of factors extracted, same items assigned to each factor) within a similar sample. A more rigorous threshold for replicability would be seeing the same number of factors extracted, the same items assigned to the same factors, and the same range of magnitudes of factor loadings (within reason). Stronger replicability gives researchers more confidence that a particular scale will behave as expected in data subsets or a new sample.

The EFA replication procedures explored in this chapter will provide readers information about the extent to which their EFAs meet these reasonable and basic expectations for replicability.

Replication or cross-validation in the literature. In the clinical literature, the use of factor scores (weighted averages of items based on factor loadings) is a contentious issue as factor loadings (and as noted in Chapter 3, even factor structure) can vary dramatically across groups, thus leading identical patient or participant responses to vary considerably across samples where factor loadings differ . Thus, for example, Floyd and Widaman (1995) suggest cross-validation procedures for factor scores, similar to those recommended for regression prediction equations. This recommendation highlights the importance of knowing how well a solution within one sample – even a very large, representative sample—generalizes.

Similarly, Briggs and Cheek (1986) argued almost three decades ago that one of the critical concerns to personality psychologists (and personality measurement) should be

---

[30] As a field, we have traditionally referred to scales as "reliable" or "unidimensional", but methodologists since Lord and Novick (1968) caution that *instruments* do not have reliability, only *scores from particular samples* do (see also Wilkinson and the Task Force on Statistical Inference, 1999). Despite this, we should have a reasonable expectation for instruments to have the same basic structure across samples if we are to have any rational basis for the science of measurement within the social sciences.

replicability of factor structure, demonstrating replicability issues within a commonly used Self-Monitoring scale.

One high-profile application of EFA replication techniques was an ambitious attempt by Costa and McRae (1997) to examine whether the commonly-held Five Factor Model of personality generalized across six different translations of their revised NEO personality inventory. In this application, strong replication across cultures and languages including English, German, Portuguese, Hebrew, Chinese, Korean, and Japanese samples not only confirmed the goodness of the translations of the instrument, but the universality of the five factor model.

What I would rather have seen, particularly in the case of Costa and McRae, was a multi-group confirmatory factor analysis, wherein they could have used inferential statistics to determine if various aspects of the factor model were significantly different across groups. While interesting, their work is yet another example of application of exploratory techniques for confirmatory purposes.

Procedural aspects of replicability analysis

For those familiar with shrinkage analyses and cross-validation of prediction equations in multiple regression, these procedures and suggestions will hopefully feel familiar. Replicability analyses in EFA (e.g., Thompson, 2004) can be conducted in two different ways: via *internal* or *external* replication. In internal replication, the researcher splits a single data set into two samples via random assignment. In external replication, the researcher uses two separately gathered datasets. In brief, replicability analysis occurs as follows:

1. EFA is conducted on each sample by extracting a fixed number of factors using a chosen extraction method (i.e., maximum likelihood or PAF) and rotation method (i.e., Oblimin or Varimax).
2. Standardized factor loadings are extracted from the appropriate results for each sample (e.g., pattern matrix if using an oblique rotation), creating a table listing each item's loading on each factor within each sample.
3. Factor loadings and structures are then compared.

Unfortunately, references on this topic do not go into depth as to how researchers should perform this comparison and what the criteria is for strong vs. weak replication, and how to summarize or quantify the results of the replication. Thus, my student at the time, David Fitzpatrick, and I developed some procedures that made sense to us (Osborne & Fitzpatrick, 2012). Hopefully, they will be sensible to you as well.

Quantifying Replicability in Exploratory Factor Analysis.

Researchers have been proposing methods of quantifying and summarizing this sort of analysis since the early 1950s. While invariance analysis in confirmatory factor analysis should be considered the gold standard for attempting to understand whether an instrument has the same factor structure across different groups (randomly constituted or otherwise), for researchers wanting to explore replication in EFA across

---

different groups, simple summary measures are to be preferred. We should leave the rigorous, statistically complex comparisons to invariance analysis and CFA.

One method of summarizing EFA replication analyses include a family of coefficients first presented by Kaiser, Hunka, and Bianchini (1971). This "similarity coefficient" utilized the cosines between the unrotated and rotated axes, but had faulty assumptions (and therefore are invalid from a mathematical point of view; see ten Berge (1996); see also Barrett (1986)) and could yield similarity coefficients that indicate strong agreement when in fact there was little agreement. Thus, they are inappropriate for this purpose.

Tucker (1951) and Wrigley and Neuhaus (1955) have presented congruence coefficients that seem less problematic (ten Berge, 1986) but are also controversial (c.f., Barrett, 1986). For example, Tucker's (1951) Congruence Coefficient examines the correlations between factor loadings for all factor pairs extracted. Yet as Barrett (1986) correctly points out, these types of correlations are insensitive to the magnitude of the factor loadings, merely reflecting the patterns. [31] For our purposes, which is to examine whether the factor structure and magnitude of the loadings are generally congruent, this insensitivity to magnitude of loadings is problematic. We prefer a more granular analysis that examines (a) whether items are assigned to the same factors in both analyses, and (b) whether the individual item factor loadings are roughly equivalent in magnitude—the former being the basic threshold for successful replication, the latter being a more reasonable, stronger definition of replication.

Assessing whether the basic factor structure replicated. Regardless of whether the researcher is performing *internal* (a single sample, randomly split) or *external* (two independently gathered samples) replication, the researcher needs to perform the same EFA procedure on both, specifying the same number of factors to be extracted, the same extraction and rotation procedures, etc. Researchers should then identify the strongest loading for each item (i.e., which factor does that item "load" on), and confirm that these are congruent across the two analyses. For example, if item #1 has the strongest loading on Factor 1, and item #2 has the strongest loading on factor #2, that pattern should be in evidence in both analyses. If any items fail this test, we would consider these analyses to fail to meet the most basic threshold of replicability: structural replicability. There is therefore little reason to expect factor structure to replicate in any basic way in future samples.

If there are a small percentage of items that seem volatile in this way, this replication analysis may provide important information—that these items might need revision or deletion. Thus, replication can also serve important exploratory and developmental purposes. If a large number of problematic items are observed, this represents an opportunity for the researcher to revise the scale substantially before releasing it into the literature, where this volatility might be problematic.

Assessing strong replication in EFA. If a scale passes the basic test of having items structurally assigned to same factors, the other important criterion for strong

---

[31] We could go on for many more pages summarizing various historical approaches to summarizing congruence. For the sake of parsimony we will simply refer the readers to the above-cited resources that give thorough coverage of the issues.

replication is confirming that the factor loadings are roughly equivalent in magnitude. We believe that because we are still in exploration mode, simple metrics serve our goal well. We advocate for simply subtracting the two standardized (rotated) factor loadings for congruent items, and squaring the difference. Squaring the difference has two benefits: eliminating non-important negative and positive values (if one loading is .75 and one is .70, subtracting the first from the second produces a -0.05, and subtracting the second from the first produces a 0.05, yet the direction of the difference is unimportant—only the magnitude is important) and highlighting larger differences. Researchers can then quickly scan the squared differences, and either confirm that all are small and unimportant, or identify which items seem to have large differences across replication analyses.

An example of replication analysis.

For this example, we return to the scale I developed to measure identification with academics, and which I had you perform EFA on the pilot data from a community college sample (Osborne, 1997). This example is from a different sample of 1908 participants from several community colleges around the USA. This published version of the SPQ is a scale of 13 questions designed to measure identification with academics (also called selective valuing or domain identification in the self-concept literature; for a recent article on this concept, see Osborne & Jones, 2011). The SPQ Scale questions relevant to this data set are listed below (measured on a scale of 1 (strongly disagree) to 5 (strongly agree). * Indicates that item is reverse coded).

Items in the School Perceptions Questionnaire (SPQ) Scale:

1. Being a good student is an important part of who I am.
2. I feel that the grades I get are an accurate reflection of my abilities.
3. My grades do not tell me anything about my academic potential.*
4. I don't really care what tests say about my intelligence.*
5. School is satisfying to me because it gives me a sense of accomplishment.
6. If the tests we take were fair, I would be doing much better in school.*
7. I am often relieved if I just pass a course.*
8. I often do my best work in school.
9. School is very boring for me, and I'm not learning what I feel is important.*
10. I put a great deal of myself into some things at school because they have special meaning or interest for me.
11. I enjoy school because it gives me a chance to learn many interesting things.
12. I feel like the things I do at school waste my time more than the things I do outside school.*
13. No test will ever change my opinion of how smart I am.*

To demonstrate this technique, we used *internal replicability analysis*, randomly splitting the original sample into two independent samples that were then analyzed separately using specific extraction and rotation guidelines based on prior analyses of the scale. In this example we report a two-factor solution (the factor structure suggested by previous

research on the scale) as well as 3- and 4-factor solutions to demonstrate how mis-specification of a factor model can quickly become evident through replication analysis.

Two-factor replication analysis. The basic overview of the replication is presented in Table 4.1. As you can see in this table, replication of this scale fails to meet the initial criterion, structural replication. Specifically, looking at the factor loadings, you can see Question 12 has the highest factor loading on Factor #2 in the first analysis and on Factor #1 in the second analysis. This item is probably not a good one, due to the cross-loading, and would benefit from revision or deletion. All other items have their strongest loading on congruent factors, so if we delete Question 12, we would say that the factor structure of the scale meets the basic level of replication. The next step is to look at the squared differences in the factor loadings. These range from 0.0000 to 0.01, indicating that the largest difference between the standardized factor loadings is |.10|-- which is not bad. We would suggest that once the squared differences achieve a magnitude of .04—indicating a difference of |.20| -- that is when a researcher may begin to consider factor loadings volatile.

Table 4.1
Two- factor SPQ replicability analysis, ML extraction, Oblimin rotation

		Sample 1			Sample 2			
	Comm	Factor Loading		Comm	Factor Loading		Squared difference
		1	2		1	2	
SPQ 01	0.42	0.66		0.36	0.60		0.0036
SPQ 02	0.29	0.51		0.29	0.51		0.0000
SPQ 03	0.26		0.41	0.23		0.39	0.0004
SPQ 04	0.33		0.52	0.36		0.57	0.0025
SPQ 05	0.44	0.64		0.48	0.71		0.0049
SPQ 06	0.28		0.54	0.19		0.44	0.0100
SPQ 07	0.12		0.35	0.14		0.38	0.0009
SPQ 08	0.26	0.52		0.31	0.58		0.0036
SPQ 09	0.39	-0.44	0.36	0.39	-0.50		0.0036
SPQ 10	0.28	0.54		0.27	0.54		0.0000
SPQ 11	0.50	0.71		0.54	0.74		0.0009
SPQ 12	0.35	-0.34	0.42	0.38	-0.45	0.31	*failed*
SPQ 13	0.15		0.40	0.22		0.49	0.0081
Eigen:		2.76	1.60		3.06	1.66	

Note: Loadings less than 0.30 were suppressed to highlight pattern. Pattern coefficients reported

Three-factor replication analysis. As mentioned above, this should replicate poorly as a 3-factor solution is not a strong solution for this scale. As you can see in Table 4.2, problems are immediately obvious. Even with such a large sample, three of the thirteen items failed to replicate basic structure—in other words, they loaded on

non-congruent factors. Further, Question 8 is problematic because it is not clear what factor to assign it to in the first analysis (it loads 0.32 on both factors 1 and 3), whereas in the second analysis it loads strongly on Factor 1, so it could be argued that three of the thirteen items failed basic structural replication. Beyond these three, the squared differences for the loadings were within reasonable range (0.0000-0.0225) except for Question 8, which had a 0.0529, reflecting a large change in factor loading from 0.32 to 0. 55. This would be a second red flag for this item, if the researcher decided to let the issue of structural replication pass.

Table 4.2
Three- factor SPQ replicability analysis, ML extraction, Oblimin rotation

	Sample 1				Sample 2				
	Comm	Factor Loadings			Comm	Factor Loadings			Squared
		1	2	3		1	2	3	Difference
SPQ 01	0.45	0.43		0.39	0.39	0.57			.0196
SPQ 02	0.57			0.70	0.47	0.45		0.41	*failed*
SPQ 03	0.36		0.32	-0.45	0.36		0.45	-0.32	*failed*
SPQ 04	0.34		0.47		0.35		0.57		.0100
SPQ 05	0.45	0.60			0.47	0.69			.0081
SPQ 06	0.30		0.55		0.18		0.43		.0144
SPQ 07	0.15		0.39		0.14		0.36		.0009
SPQ 08	0.27	0.32		0.32	0.34	0.55			.0529
SPQ 09	0.39	-0.39	0.36		0.45	-0.54			.0225
SPQ 10	0.31	0.57			0.27	0.54			.0009
SPQ 11	0.60	0.76			0.56	0.76			.0000
SPQ 12	0.38	-0.37	0.45		0.52	-0.51	0.32	0.31	*failed*
SPQ 13	0.16		0.35		0.21		0.48		.0169
Eigen:		2.45	1.46	1.84		3.09	1.69	0.58	

Note: Loadings less than 0.30 were suppressed to highlight pattern. Pattern coefficients reported.

Four-factor replication analysis. I decided not to show the replication table for this analysis as the basic structural replication failed dramatically – and unsurprisingly— with ten of the thirteen items loading on non-congruent factors. Of the other three, one changes from 0.99 to -0.58, which represents a massive shift in magnitude, another shifts from -0.52 to 0.33, again a relatively large shift, and the final one shifts modestly from 0.44 to 0.37. In almost every way, this analysis demonstrates everything that can go wrong with a replication analysis, and as such, does not require a full-page table to describe. If you are curious about how bad this replication was, perform this replication as an exercise at the end of this chapter, and then hope you never see a replication table like it with your own data!

Appropriately large samples make a difference. In Table 4.3, I replicate the two-factor analysis presented in Table 4.1 but with two random samples of N=100 each, much smaller than the almost N=1000 samples in Table 4.1. In this analysis, you

can see two of the thirteen items loaded on non-concordant factors (interestingly, not the originally-troublesome Question 12), and two more items had troublingly large differences in factor loadings. Question 1 loaded 0.77 in the first analysis and 0.56 in the second analysis. As you can see from the communality estimates, that led to a large decrease in the communality for this item—and a squared difference of over 0.04. Additionally, Question 7 had a loading of 0.82 in the first analysis and 0.39 in the second analysis, again leading to a large change in communality and a squared difference of 0.1849. Thus, even if a researcher deleted the two troublesome items, two others showed non-replication of magnitude of factor loading. As previous authors have noted, EFA is a large-sample procedure, and replications with relatively small samples may lead to more volatility than one would see with larger samples. With over 900 in each sample, this scale looks relatively replicable, but with only 100 in each sample there are some serious questions about replicability.

Table 4.3
Two- Factor SPQ Replicability Analysis, ML Extraction, Oblimin Rotation; Small Samples

	Sample 1			Sample 2			
	Comm	Factor Load		Comm	Factor Load		Squared
		1	2		1	2	difference
SPQ 01	0.55	0.77		0.31	0.56		.0441
SPQ 02	0.42	0.62		0.39	0.57		.0025
SPQ 03	0.29		0.52	0.35		0.57	.0025
SPQ 04	0.27	-0.34		0.35		0.58	*failed*
SPQ 05	0.56	0.68		0.37	0.62		.0036
SPQ 06	0.32		0.56	0.30		0.55	.0001
SPQ 07	0.62		0.82	0.15		0.39	*failed*
SPQ 08	0.34	0.61		0.33	0.56		.0025
SPQ 09	0.40	-0.49		0.40		0.46	*failed*
SPQ 10	0.21	0.46		0.32	0.58		.0144
SPQ 11	0.46	0.64		0.49	0.71		.0049
SPQ 12	0.50		0.46	0.24		0.34	.0144
SPQ 13	0.19		0.40	0.35		0.60	.0400
Eigen:		2.76	1.60		3.06	1.66	

Note: Loadings less than 0.30 were suppressed to highlight pattern. Pattern coefficients reported

Examining the communalities, you can also easily see that these statistics seem to vary widely across the two samples, some almost double- or half- that of the comparable communality in the first analysis. One, (SPQ 07) decreased from 0.62 to 0.15, less than a quarter of the first communality.

It is also useful to point that merely deleting items that are troublesome in this analysis may not be ideal. A researcher performing the analyses in Table 4.3 first (with small samples) would delete two items that showed fine replicability in Table 4.1 (larger samples), and would retain the one troublesome item. Thus, researchers should ensure

they have large, generalizable samples prior to performing any exploratory factor analysis.

Chapter 4 Summary: Is replication important in EFA?

Although authors have been presenting methods for summarizing replication in EFA for half a century and more, most summarization techniques have been flawed and/or less informative than ideal. In the 21st century, with CFA invariance analysis as the gold standard for assessing generalizability and replicability, replication within EFA has an important role to play—but a different role than half a century ago. Today, replication in EFA is a starting point, -- it adds value to EFA analyses in that it helps indicate the extent to which these models are likely to generalize to the next data set, and also in helping to further identify volatile or problematic items. This information is potentially helpful in the process of developing and validating an instrument, as well as for potential users of an instrument that has yet to undergo CFA invariance analysis.

However, there are often barriers to replication analysis. Foremost amongst these barriers is the lack of adequate sample size in most EFAs reported in the literature. The first priority for researchers should be adequate samples. The second should be estimation of the replicability (or stability) of the model presented. In the next chapter I review bootstrap analysis as a potential solution to this issue, as it allows use of a single, appropriately large sample to estimate the potential volatility of a scale.

Chapter 4 Exercises

1. Download the SPQ data from the book website and split the file into two randomly chosen samples. Repeat the EFA and replication as performed in the chapter to see if you get similar results.
    a. After performing a basic replication with the full sample, randomly select two smaller samples from the large sample and see how that influences replicability of EFA results.
    b. As recommended above, require four factors be extracted and then perform a replication to see how amazingly poor that analysis went.
2. Return to the engineering data from Chapter 2. Test whether those EFA results replicate by randomly splitting the file into two samples. With an original N = 372, the samples will be smaller, and thus more volatile. Will the strong factor structure previously observed be maintained in two smaller samples?
3. Select two small random samples from the Marsh SDQ data (N=100 each). Replicate the EFA from Chapter 2 (ML extraction, Promax rotation, extracting 3 factors) and compare your results to those from the very large sample reported in that previous chapter. Then compare the results from the two samples to each other as we did in these replication examples. Does the factor structure replicate when using such a small sample?
    a. Once you have done that, select two large samples from the Marsh data (N=1000 each) and see if those two samples replicate better.

5 BOOTSTRAP APPLICATIONS IN EFA

Resampling analyses are relatively new to me (I first used them in preparing my last book on logistic regression) but I have been impressed by the possibilities of bootstrap in particular to answer questions that we never could ask before. For example, given a particular sample, we can ask how likely it is that our findings will replicate, or how broad our confidence intervals are around a particular statistic. This is new, and relies entirely upon high powered computers that currently sit on our desks. It is a limited methodology, in that it cannot do everything that many people think it can, but it might have a place in helping us understand how strong an EFA analysis really is, particularly if you do not want to split a sample in half to do the type of replication analysis performed in the previous chapter.

Some background on resampling

Wilkinson and APA (1999) crafted a vision of modern quantitative methodology that moved away from (or complimented) our historical reliance upon null hypothesis statistical testing (NHST) to a more nuanced approach of quantitative reasoning that involved effect sizes, confidence intervals, and of course, confidence intervals for effect sizes. Leading thinkers in quantitative methods (to name just a few: Bruce Thompson Geoff Cumming, and Fiona Fidler) have been attempting to move the field of quantitative methods forward to more nuanced thinking about effect sizes and confidence intervals. Peter Killeen made a recent suggestion that we could calculate the probability that a finding would replicate in order to inform readers of the potential utility of a given set of results. Jacob Cohen's long crusade to bring power into consideration is in a similar vein of wanting to have researchers be more thoughtful and informed about their data.

However, while effect sizes and confidence intervals and even power has been increasingly evident in some literatures it is beyond the grasp of most researchers to produce confidence intervals for commonly-reported and important statistics like effect sizes (e.g., eta-squared), reliability estimates (e.g., Cronbach's alpha), or widely reported exploratory techniques like exploratory factor analysis.

It is not routine for researchers to report any confidence intervals relating to EFA, and indeed, it is not routine for replication to be considered much at all. However, I hope the appeal of this is intuitive. If I perform an EFA and then calculate 95% confidence intervals for the relevant statistics, it helps a reader understand how precise my estimates might be. Very broad CIs might signal to the reader that the EFA is not very precise, and therefore, not terribly informative. Narrow CIs, on the other hand, might signal to the reader that the analysis is worth considering seriously. It does not necessarily mean that the EFA reflects the population parameters exactly, but it is more likely to be of use than one with low precision (broad CIs). This type of analysis, while not routine, is simple through bootstrap resampling methodologies (DiCiccio & Efron, 1996; Efron & Tibshirani, 1994). Thompson (1993) argued that bootstrap analyses can provide inferences about the potential replicability of a result as well as empirical confidence intervals that can also provide alternative and complimentary information about whether an effect is significant (i.e., different from 0, for example).

A central hallmark of science is replication, the stability or replicability of an effect is also important (Killeen, 2008; Thompson, 2002; Yu, 2003). While not perfect, resampling methods can inform the researcher as to the relative stability or instability (i.e., replicability or non-replicability) of an effect or result.

What is bootstrap resampling analysis?

Bootstrap resampling is a methodology that is increasingly popular now that desktop computers can perform thousands of analyses per second, even with large samples. There are many good references on bootstrap and other resampling techniques. The brief overview here is not meant to be exhaustive, but rather to give enough information for you to understand the rest of the sample.

The origin of bootstrap analysis was seeking a solution for researchers with inadequate samples. Bootstrap resampling takes an existing sample (say, of 50 participants) and randomly selects (with replacement) a certain number of related samples of N=50 based on those original 50 subjects.[32] The procedure is called "resampling" because it treats the original sample as fodder for an unlimited number of new samples. By resampling with replacement, we can get 3 copies of the 14th person in the sample, none of the 15th, and one copy of the 16th person. Perhaps in the next sample there will be one copy of both the 14th and 15th persons, but none of the 16th. Thus, the samples are related, in that they all derive from the same master sample, but they are not exactly the same as each individual can be present in varying degrees or not in each resampling.

The goal of this resampling methodology is to provide a large number of permutations of the sample, and then to analyze all the samples and provide summary statistics (average effect, 95% confidence intervals, etc.) for those effects. As Thompson (2004) argued, bootstrap resampling analysis can be conceptualized as one method for estimating the sampling distribution of the relevant statistic from the population. Most scholars familiar with bootstrap resampling will agree with what I have said thus far, but likely will stop agreeing at this point. There are a wide number

---

[32] As far as I know, bootstrap resampling always uses the same sample size as the original sample

of opinions on what bootstrap resampling is good for, and what it is not good for. You will get my opinion on that in this chapter (hopefully with sufficient empirical evidence to make my case) but be aware that there are strong passions around this issue (much like principal components analysis…).

What can bootstrap resampling do, and what should it not be used for?

Many early adopters of the procedure saw bootstrap analysis as a panacea for small or biased samples, reasoning that with enough resampled data sets, the bias and small sample would be compensated for, and would provide a better estimate of the population parameters than the original sample by itself. My experiments with bootstrapping of small, biased samples indicates that the samples tend not to be self-correcting. In other words, bootstrapping a small, biased sample tends to lead to promulgating that bias. Large biased samples are probably in the same category. In my previous book on logistic regression, I tested some of these assertions, finding that fatally biased samples do not tend to do anything other than produce biased results, even with many thousands of bootstrap analyses.

Bootstrap can also not do much to help small samples. To be sure, you can endlessly resample the same small sample, but there is limited information in the small sample. One cannot build something out of nothing. The best one can do with a small sample is to bootstrap some confidence intervals and evaluate just how imprecise the parameter estimates are.

Bootstrap analyses can help identify when there are inappropriately influential data points in a sample. If one does thousands of resampling analyses, and they are distributed with a skew, the long tail is likely due to the influence of a few cases. However, there are easier ways to detect inappropriately influential data points, and in those bootstrap explorations I did with logistic regression, I was able to show that cleaning the data *prior to the bootstrap analysis* often yielded much better results. Thus, if you have a sample you are intending to bootstrap, it is best to do some preliminary data cleaning first.

Thus, bootstrap can be a valuable tool in the statistician's toolbox, but it is not a panacea for all our ills. It cannot fix a fatally flawed sample, it cannot compensate for an inappropriately small sample, and it is not best used to determine if there are influential cases in the data. But given a reasonable sample, bootstrap resampling can do some interesting things. It can provide confidence intervals for things like effect sizes that we really cannot get any other way. It can provide information on the precision of the results, and it can give some information in a single sample that is helpful in determining whether a solution will replicate or not. In other words, if one performs an appropriate bootstrap analysis of a reasonable sample, and one sees relatively narrow confidence intervals, one can say that the solution arrived at is more precise than if one has very broad confidence intervals. Further, if those confidence intervals are narrow and precise, it is likely that a similar sample will produce similar results. If the confidence intervals are wide and sloppy, it is not likely that a similar sample would produce similar results.

A simple bootstrap example in ANOVA.

Let us take a simple (non-EFA) digression to explore an example from a collaboration with a colleague Philip Gabel. Dr. Gabel contacted me exploring a new physical therapy methodology for rehabilitating knee injuries, specifically whether this new technique provided more exercising of the quadriceps, the muscles in the thigh that help keep the knee in alignment.[33] Gathering data is time-consuming and onerous, so he sent me a small pilot sample of 21 patients to see if this line of research was worth pursuing (and might ultimately lead him to a publication). Obviously a sample of 21 is not large, and when there are four different techniques being compared to this fifth technique (called slacklining; tested in a repeated measures format) is interesting but not large enough (in my opinion) to run to a top tier journal with. So my task was to assess whether there is promise in the data. To do this, I performed an initial repeated measures ANOVA to see if his hunch was correct. The first four techniques are standard physical therapy techniques, and the fifth is the experimental slacklining technique. The dependent variable is an electrical measure of muscle activation, EMG. As you can see in Table 5.1, below, not only did the fifth condition stand out as substantially different, the 95% confidence intervals for this group barely overlapped with the other four. This is a good indicator that there is a strong effect here.

Table 5.1
EMG measure in knee rehabilitation patients across five activities

EMG	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	82.714	14.935	51.561	113.868
2	91.143	15.915	57.945	124.341
3	97.095	13.464	69.010	125.181
4	90.571	13.934	61.506	119.637
5	141.381	17.531	104.812	177.950

To answer my colleague's question, I performed bootstrap analysis with B=2000 sample replications, each of N=21. The central questions were: (a) what is the expected range of effect sizes and their expected stability, and (b) what percent of replicate samples produce significant effects. Because of the small size, I report Greenhouse-Geisser corrected statistics rather than assuming homogeneity of variance across conditions. Using a macro in SPSS, which I will share with you later on (and make available on the web site), I performed these 2000 analyses of the resampled data, and then summarized the bootstrap analyses for some key statistics: F of the effect, *p* value, and partial eta-squared (an indicator of percent variance accounted for, which I often use as an effect size in ANOVA analyses). Of course, it is not easy or routine to get 95% confidence intervals for *F*, *p* values, or eta-squared statistics. Using the macro in SPSS, I could have retrieved almost any statistic output in the repeated measures

---

[33] I just vastly exceeded my knowledge of knee anatomy, so forgive me if I mis-stated something. The gist of the thing is most important.

procedure.[34]

Table 5.2
Summary of bootstrap analysis of EMG data.

	F	$p <$	Significant percent	Partial Eta-Squared
Mean	13.78	.00054	99.8%	.40
Median	13.03	.00002	100.0	.39
Std. Deviation	5.12	.0033	3.9%	.08
Minimum	2.44	.000	0.0%	.11
Maximum	56.99	.096	100.0%	.74
2.5%	5.95	.00	100.0%	.23
97.5%	25.48	.0043	100.0%	.56

As Table 5.2 shows, the average F from 2000 samples was 13.78, and 95% of the samples produced Fs ranging from 5.95 to over 25. Not surprisingly, the 95% confidence interval for p was 0.00001 to 0.0043, all well below 0.05. The third column, significant, was recoded into 0 if p was .05 or greater and 1 if p fell below the magical .05 level. As you can see, 99.85% of the 2000 samples were significant at $p < .05$, leaving a strong expectation that future analyses would replicate a significant result. Finally, the effect size (eta squared) averaged 0.396, with a 95%CI of 0.229 to 0.560. Since this 95% CI for effect size does not come close to zero (in fact, the lowest of the 2000 estimates was 0.11, not an insignificant effect size), there is good reason to expect that a similar sample would produce similar results. Further, the precision of the estimates was strong. For example, the standard deviation of the sampling distribution for eta-squared is 0.084.

So why are we talking ANOVA in a book on factor analysis? This is a clear example of the power of bootstrapping. Based on these analyses, I could make a reasonable expectation that my colleague was on the right track, that another larger sample was likely to be significant at $p < .05$, and that there was likely to be a rather impressive effect size. The estimate for eta-squared was an average of 0.39, and although the estimate was not precise (i.e., it had a large 95% CI around it), the expected range was reasonable enough to predict replication of a worthwhile effect size. Had the results of the bootstrap been different (many non-significant effects, the 95% CI for eta squared including 0 or very small effects) then we would not have had confidence of replication (or finding similar results in an expanded sample).

Did bootstrap analyses inform potential replication? Ultimately, my colleague collected 52 cases in his study, and it is currently in press. I went back and examined only the 34 cases that he gathered after the initial bootstrap analysis to see if those projections were useful. The F in the new sample was 13.15, $p <. 0001$, with an eta-

---

[34] If you pay for the bootstrap module in SPSS, it is even easier. I believe SAS routinely comes with bootstrap capabilities, and R seems to have excellent capabilities as well in this area.

squared of 0.32. Not only were the statistics for the new sample well within the 95% confidence intervals, they were close to the median or mean of the bootstrapped samples. Thus, at least in this case, the bootstrap analysis of a small sample was informative about the next group of subjects and the replicability of the initial analysis.

Confidence intervals for statistics in EFA

Let us turn back to the task at hand. We can bootstrap an exploratory factor analysis[35] to obtain similar information to the replication analysis in the previous chapter—if we are strategic about it. EFA does not have significance tests nor easily interpretable effect sizes, but there are many things we can bootstrap in order to get some information about potential robustness of the results. However, be forewarned that EFA is a complex procedure with lots of effects. Unlike ANOVA, which had only a few we bootstrapped, EFA can have many communalities, eigenvalues, and factor loadings. So while it does not take a terribly long time with modern computers (it took my computer about 5 minutes to run 2000 bootstrap analyses and save them to a data file—the process that takes the longest is examining all the different parameters one wants to examine. The time to perform analyses seems to expand dramatically moving to 5000 analyses).

I will add the SPSS syntax I use to bootstrap at the end of the chapter and also make it available via the book's web site. It is a bit complex, but there are only a few aspects of the macro one needs to adjust, and you should be able to use these macros right out of the box with only slight modification, which I will try to note clearly. Of course, if you purchase the SPSS bootstrap module, that makes life easier, and SAS and R also have incorporated easier ways to do bootstrapping automatically. But I have been an SPSS user for many years, and derive some satisfaction out of having this level of granular control over the process.

I will also make the bootstrap analyses data sets available so you can replicate the results and play with the bootstrap findings without performing the actual bootstrapping if you like. Remember, if you start from scratch and perform bootstrap analysis, it is likely your results will be similar to mine, but they might not be exactly identical. If you play with my data sets, you should see similar results.

Bootstrap example 1: Engineering data

The analyses in Chapter 2 indicated that this data had two small, clear factors. The sample was relatively small, however, so one question we could ask is whether it is likely the two-factor model would replicate. With such a small sample, I would be hesitant to split it and perform replication analyses as in Chapter 4. Thus, it is possible

---

[35] After writing this section I discovered that Bruce Thompson has addressed many of these ideas and issues in his 2004 book on exploratory and confirmatory factor analysis (Chapter 9). We should not be surprised, as Dr. Thompson has routinely been far ahead of the field. It is possible this chapter gave me some of the ideas for this chapter, although I do not remember reading it prior to writing. Where there is confusion as to whether Thompson should get credit or I should get credit, please assign it to Thompson.

bootstrap analysis will be informative. Other questions could involve estimation of reasonable ranges of factor loadings and communalities for each of the variables. Let's take things one step at a time. To perform this bootstrap analysis, I used the same settings as in Chapter 2 with respect to extraction and rotation.

Eigenvalues for first three factors initially extracted. We expect two factors to be extracted (by eigenvalue > 1 criteria, which is simplest in this analysis). Thus, if we examine the first three extracted, and examine whether any of the bootstrap analyses produced a third factor with initial eigenvalue greater than 1, we can explore the likelihood that this basic issue would replicate in a similar sample. A histogram for the 5000 eigenvalues extracted first is below in Figure 5.1, so you can see how bootstrap analyses work.
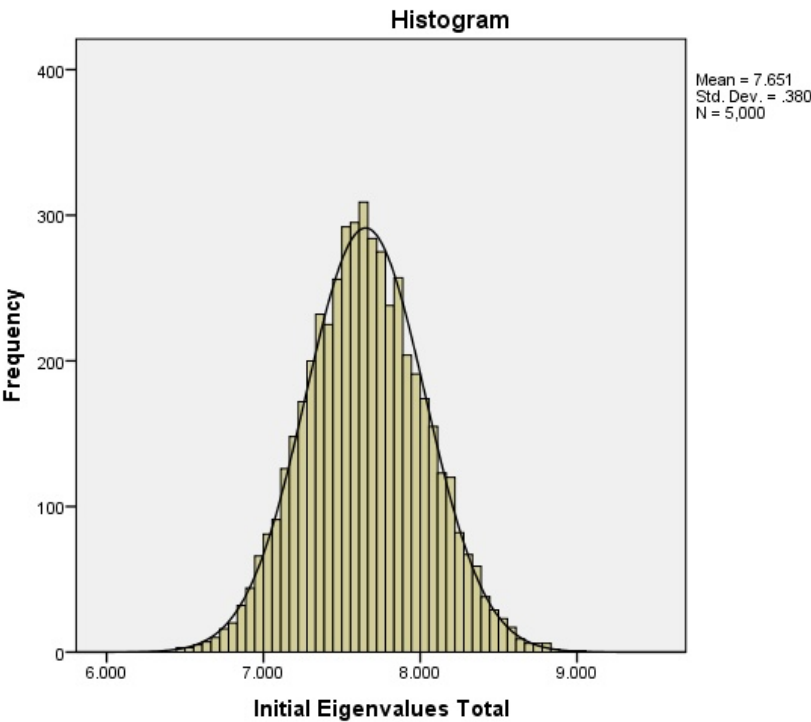
Figure 5.1: Distribution of first eigenvalue extracted over 5000 bootstrap analyses. Mean is 7.65, ranging from 6.47 to 9.03, with a 95% CI of 6.93, 8.40.

Table 5.3a
Bootstrap results for first three eigenvalues extracted

	Mean Eigen	Std. Dev	95% CI	Mean % variance	Std. Dev	95% CI
Factor 1	7.65	0.38	6.93, 8.40	54.65%	2.72	49.48, 59.99
Factor 2	3.50	0.27	2.97, 4.03	25.02%	1.93	21.20, 28.79
Factor 3	0.49	0.05	0.39, 0.60	3.49%	0.39	2.78, 4.27

The first three eigenvalues (presented in Table 5.3a) are congruent with our initial analysis in Chapter 2. Furthermore, the third eigenvalue does not rise close to 1.0 in any bootstrapped data set. The two-factor solution is therefore strongly supported and supports a reasonable expectation that this factor structure would be found in another similar data set.

Communality results. In Table 5.3b I present the bootstrap statistics for communalities and factor loadings. The communalities are relatively strong, consistently, ranging from 0.68 to 0.82, with 95% confidence intervals that are reasonably narrow also. This leads us to expect that another replication in a similar sample would generally extract relatively strong communalities.

Table 5.3b
Bootstrap results for communalities and factor loadings

	Communalities		Bootstrapped Pattern Coefficients		Original Pattern Coefficients	
	Bootstr.	Orig.	Factor 1	Factor 2	Factor 1	Factor 2
EngProb1	.75 (.69, .81)	.742	.84 (.79, .89)	.02 (.00, .06)	.859	-.016
EngProb2	.71 (.64, .77)	.695	.83 (.76, .89)	.07 (.01, .13)	.841	-.071
EngProb3	.76 (.70, .82)	.752	.91 (.87, .94)	.03 (.00, .07)	.879	-.008
EngProb4	.80 (.74, .85)	.792	.91 (.87, .94)	.03 (.00, .07)	.909	-.025
EngProb5	.80 (.74, .85)	.790	.89 (.85, .93)	.03 (.00, .07)	.886	.021
EngProb6	.78 (.71, .84)	.766	.87 (.82, .91)	.03 (.00, .07)	.869	.020
EngProb7	.80 (.74, .85)	.786	.87 (.82, .91)	.03 (.00, .09)	.868	.033
EngProb8	.68 (.62, .75)	.666	.79 (.73, .84)	.07 (.01, .14)	.790	.072
INTeng1	.70 (.59, .77)	.674	.05 (.00, .12)	.79 (.71, .87)	.042	.801
INTeng2	.82 (.74, .88)	.802	.03 (.00, .07)	.92 (.87, .96)	-.023	.921
INTeng3	.82 (.77, .87)	.816	.02 (.00, .05)	.92 (.89, .95)	-.014	.922
INTeng4	.82 (.70, .90)	.806	.02 (.00, .05)	.91 (.83, .97)	-.001	.904
INTeng5	.79 (.72, .85)	.781	.02 (.00, .06)	.88 (.83, .92)	-.007	.897
INTeng6	.75 (.67, .82)	.739	.02 (.00, .06)	.86 (.80, .91)	.009	.864

Note: 95% CIs in parentheses. Factor loadings highlighted are those expected to load on the factor.

Caution about bootstrapping factor loadings: The order in which the factors

are extracted can be arbitrary. In some of the bootstrapped analyses, interest was extracted as the first factor, and in some, the problem solving factor was extracted first. This can present a problem for bootstrapping as the data file does not know which was extracted first. Thus, in my syntax file I have an algorithm to identify cases where problem solving was extracted second and swapped factor loadings to ensure the results of the bootstrap are appropriate. Another complication is that factor loadings can be negative or positive, generally arbitrarily,[36] and thus I convert all values to absolute values to eliminate the possibility of one data set having a loading of -0.88 and another with +0.88, which means the same, generally but does not do good things for bootstrap analysis. In my data set, @1 and @2 are the variable names for the factor loadings for the first and second factors, respectively, and I compare the absolute value of the loadings on factor 1 and factor 2.[37]

```
String Factor (A2).
Compute factor= substr(var1, 1, 2).
execute.
compute F1=abs(@1).
compute F2=abs(@2).
execute.
do if (Factor = "En").
if (abs(@2)>abs(@1)) F1=abs(@2) .
if (abs(@2)>abs(@1)) F2=abs(@1) .
end if.
execute.
do if (factor = "IN").
if (abs(@1)>abs(@2)) F1=abs(@2) .
if (abs(@1)>abs(@2)) F2=abs(@1) .
end if.
execute.
```

Factor loading results. Once this is done, we can analyze the data to get 95%CIs for the factor loadings, which are also presented in Table 5.3b. As you can see, the loadings are strong on the factors they should load on (e.g., for the first factor the average bootstrapped loadings ranged from 0.79 to 0.91, with relatively narrow confidence intervals around each loading). Again, these results leading to conclusion that a replication of this original EFA analysis would be likely to produce a strong, clear factor structure similar to what we saw in Chapter 2.

---

[36] I would recommend recoding all variables to be coded positively before analysis to minimize issues like this. In some EFA analyses, factors can be arbitrarily flipped in space, leaving all loadings to be negative, but without substantively changing the meaning of the factor (if all loadings are negative or positive, but the same general magnitude, that shouldn't change the meaning as it is a constant transform). .

[37] There are further complexities to this process that leave it difficult when less clear factor structures are being extracted. If one is not careful, through this process you can artificially move the highest loadings to where you want them regardless of whether the analysis meant for them to be there. Be careful when manipulating factor loadings in this way!

Bootstrap example 2: Marsh SDQ data

In this example we will not merely replicate the strong, clear results presented in Chapter 2. Instead, we will assert that analysis of approximately 16,000 students represents the "gold standard" or population factor structure, and we will extract a small sample and see if bootstrapping a relatively small sample can accurately lead us to infer the "population" parameters with reasonable accuracy. A random sample 300 cases was selected and then subjected to bootstrap resampling and analysis (with identical extraction/rotation methods as in Chapter 2).

Imagining this small sample was your only information about this scale, I started by performing an EFA on this sample only. Three factors exceeded an eigenvalue of 1.0, which was corroborated by MAP criteria analysis, which produced a recommendation to extract three factors. Thus, if this were my only sample, theory, Kaiser criterion, and MAP criteria would lead me to extract three factors (the factor loadings for this analysis, in addition to the analysis of the entire sample of over 15,000 participants is presented below in Table 5.4b).

Bootstrap analysis of a small sample. The bootstrap analyses extracted three factors. First we will examine the basics of whether the known factor structure is likely to be replicated with such a small sample. As you can see from Table 5.4a, the "population" factor structure is largely replicated by bootstrapping this relatively small sample.

Table 5.4a

Bootstrap results for first four eigenvalues extracted Marsh SDQ

	Mean eigenvalue	Stnd. Dev.	95% CI	Mean % variance	Stnd. Dev.	95% CI
Factor 1	4.24 (4.08)	0.25	3.76, 4.77	32.63	1.96	28.89, 36.68
Factor 2	2.64 (2.56)	0.16	2.31, 2.96	20.29	1.27	17.79, 22.77
Factor 3	2.11 (2.21)	0.16	1.80, 2.43	16.26	1.21	13.83, 18.66

Note: Actual eigenvalues from full "population" in parentheses.

With three factors, the machinations to move the factors to the right column becomes more complex because there are more ways in which the factors can be re-ordered. There are also possibilities with less well-structured factors that one could have high cross-loadings that could get erroneously moved into a column it should not get moved into. Thus, if you are going to be engaging in this type of analysis, you must be very cautious and thorough in examining your data to ensure you are not mis-aligning the factor loadings, and inappropriately setting the analysis up to look more favorable than it should be. At the end of the chapter is an example of the type of syntax I used to move the absolute values around to align them.

Once the factor loadings were converted to absolute values and aligned consistently, 5000 bootstrap replications let to estimates that were not far off of the full "population" factor loadings. In this case, at least, bootstrap analyses provides a good

estimate of the population parameters.

Table 5.4b

Bootstrap results for factor loadings

Var:	Bootstrapped "small sample" Factor Loadings			"Population" Factor Loadings			Sample (N=300) Factor Loadings		
	1	2	3	1	2	3	1	2	3
Math1	.91 (.82-.97)	.04 (.00-.10)	.03 (.00-.08)	.901	-.040	-.037	.916	-.052	-.006
Math2	.87 (.80-.94)	.04 (.00-.08)	.02 (.00-.07)	.863	.008	.012	.875	.001	-.034
Math3	.87 (.82-.92)	.04 (.00-.09)	.02 (.00-.08)	.881	-.001	.023	.875	.038	-.010
Math4	.51 (.38-.63)	.09 (.01-.18)	.06 (.00-.17)	-.601	-.049	.021	-.508	-.030	-.083
Par1	.09 (.01-.18)	.75 (.63-.84)	.08 (.00-.17)	.002	.718	.023	.055	.753	-.109
Par2	.06 (.00-.13)	.63 (.50-.76)	.03 (.00-.11)	.060	-.680	.052	.043	-.626	-.003
Par3	.05 (.00-.12)	.87 (.76-.95)	.04 (.00-.12)	.028	.827	-.002	-.034	.873	.049
Par4	.09 (.01-.21)	.59 (.46-.73)	.07 (.00-.20)	-.036	-.583	-.100	.023	-.592	-.113
Par5	.07 (.01-.16)	.74 (.63-.83)	.04 (.00-.11)	.018	.749	-.030	.064	.742	-.023
Eng1	.04 (.00-.11)	.05 (.00-.11)	.77 (.69-.85)	-.005	.031	.779	.033	.035	.770
Eng2	.05 (.00-.14)	.06 (.00-.15)	.86 (.80-.91)	-.016	-.082	.842	.025	-.082	.859
Eng3	.03 (.00-.10)	.05 (.00-.11)	.85 (.78-.91)	.052	-.017	.845	.052	.000	.855
Eng4	.08 (.01-.18)	.09 (.01-.18)	.67 (.56-.77)	.060	-.102	-.609	.088	-.073	-.675

Note: 95% CIs in parentheses. Factor loadings highlighted are those expected to load on the factor. Pattern coefficients reported.

Even when the factor loadings from the bootstrap analysis are a bit off (e.g., Math4, English4), the population parameters are within the 95%CIs, reinforcing the fact that a reasonable sample, appropriately bootstrapped, can be helpful in understanding the likely values of the population parameters. Not only are the point estimates on par with what we would expect, but the precision seems to be strong also- in other words, the confidence intervals are relatively narrow for most of the effects being examined.

Chapter 5 Summary

Replication is an important principle of science. However, replication for exploratory factor analysis is tricky. Even with very large samples, the factor structure and parameter estimates are often unstable or inaccurate reflections of the population. In chapter 4 we explored a more classic replication methodology of splitting an existing sample (internal replication) or gathering two independent samples (external replication). However, given what we know of EFA and sample size, it might be simple to argue that it is always better to have a larger sample than a smaller sample. So in this chapter we explored a methodology that is relatively new to most of us- bootstrap resampling. To my knowledge, this technique has not been applied to EFA, perhaps for obvious reasons: it is very difficult to bootstrap and then evaluate all the myriad parameters produced in an exploratory factor analysis. The two examples presented in this chapter took many hours of work to produce- but if this is an important study for you, those hours are a worthwhile investment, in my opinion.

We rarely have the ability to know the "true" factor structure in the population (leaving aside the fact that factor structures can vary across subpopulations), and so most of the time we only have a sample, and a hope that the factor structure will match our theoretical model(s), and that they will generalize to new samples and studies. The past century or so of exploratory factor analysis has been almost entirely devoted to truly exploring data in an atheoretical way, or seeking to confirm that an instrument matches a theoretical model. It has been rare to see any attention given to what should be at least as important—whether the findings will generalize. In this chapter, we describe and explore a methodology to move in that direction. It is not a perfect methodology, as a poor sample will lead to bootstrap analyses that are poor- but at the least, we can show how precise our results are and how confident we can be about those estimates. As Bruce Thompson has suggested, bootstrap methods can give us valuable information about our results. With it being relatively accessible through adaptation of SPSS macros (like the ones I share in the chapter appendix and on this book's website), or use of a statistical computing package that includes bootstrapping more well-integrated, you as a researcher can use your data to provide more valuable information about your results than you might otherwise get.

You will note that I did not provide a bootstrap example of the Example 3 (Geriatric Depression Scale) data. There are several reasons for this. First, I am not sure it is valuable to bootstrap a model that is so unclear. I suspect that we would have learned that the CIs are very wide (in other words, that our point estimates are imprecise). Second, after the analyses presented in Chapter 2, I am not sure what factor structure I would test in bootstrap analyses. With more than 30 items, a bootstrap analysis would have been even more work than these, and if we looked at a 5- or 8- factor model, the processing would have been exponentially more complicated. If this was my dissertation, and I was passionate about the GDS, I would consider it anyway. But with two other good examples, I felt at liberty to pass it by. Feel free to try it on your own and let me know how it turns out.

Chapter 5 Exercises

1. Download bootstrap data sets from the examples in this chapter, as well as the syntax files (macros) for performing bootstrap replications. See if you can replicate the results (your resampled bootstrap samples will be different from mine, so you will not exactly replicate the bootstrap analyses. They should be close approximations, however.
2. Download the data files containing my 5000 resampling analyses, and explore extracting the summary statistics from them (mean, standard deviations, 95% confidence intervals, etc.)
3. With a new data set of your own, attempt a bootstrap analysis from beginning to end. I will provide some new examples you can use for exploration as I have time. Check my web site to see if new ones are up.

SPSS Bootstrap macro (download available on book website)

Example 1: Engineering data

```
DEFINE EFA_bootstrap (samples=!CMDEND)
COMPUTE dummyvar=1.
AGGREGATE
  /OUTFILE=* MODE=ADDVARIABLES
  /BREAK=dummyvar
  /filesize=N.
!DO !other=1 !TO !samples
SET SEED RANDOM.
WEIGHT OFF.
FILTER OFF.
DO IF $casenum=1.
- COMPUTE #samplesize=filesize.
- COMPUTE #filesize=filesize.
END IF.
DO IF (#samplesize>0 and #filesize>0).
- COMPUTE sampleWeight=rv.binom(#samplesize, 1/#filesize).
- COMPUTE #samplesize=#samplesize-sampleWeight.
- COMPUTE #filesize=#filesize-1.
ELSE.
- COMPUTE sampleWeight=0.
END IF.
WEIGHT BY sampleWeight.
FILTER BY sampleWeight.

****************************************.
***
**** insert syntax for EFA here
***
****************************************.
FACTOR
  /VARIABLES EngProbSolv1 EngProbSolv2 EngProbSolv3
  EngProbSolv4 EngProbSolv5 EngProbSolv6 EngProbSolv7
  EngProbSolv8 INTERESTeng1 INTERESTeng2 INTERESTeng3
  INTERESTeng4 INTERESTeng5 INTERESTeng6
  /MISSING LISTWISE
  /ANALYSIS EngProbSolv1 EngProbSolv2 EngProbSolv3
  EngProbSolv4 EngProbSolv5 EngProbSolv6 EngProbSolv7
  EngProbSolv8 INTERESTeng1 INTERESTeng2 INTERESTeng3
  INTERESTeng4 INTERESTeng5 INTERESTeng6
  /PRINT INITIAL ROTATION
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION ML
  /CRITERIA ITERATE(25) DELTA(0)
  /ROTATION OBLIMIN.
!DOEND
!ENDDEFINE.
```

> Change highlighted syntax to change details of EFA analysis to be bootstrapped

```
****************************************.
***
*** select data and run bootstrap N=5000;
*** must change GET FILE syntax to point to your data file
***
****************************************.

PRESERVE.
SET TVARS NAMES.
DATASET DECLARE bootstrap_EFA1.
OMS /DESTINATION VIEWER=NO /TAG='suppressall'.
OMS
  /SELECT TABLES
  /IF COMMANDS=['Factor Analysis'] SUBTYPES=['Total Variance
  Explained'  'Communalities'  ' Rotated Factor Matrix']
  /DESTINATION FORMAT=SAV OUTFILE='bootstrap_EFA1'
  /TAG='alpha_coeff'.

GET
  FILE='C:\Users\jwosbo04\dropbox\Public\ECPY740\Ex1_EFA\data
  .sav'.
DATASET NAME bootstrap_EFA1 WINDOW=FRONT.
Set MITERATE 10000.
Execute.
EFA_bootstrap
  samples=5000 .
OMSEND.
RESTORE.

************************************************************
***
**** allows you to select individual parameters to examine
***
************************************************************.

DATASET ACTIVATE bootstrap_EFA1.
TEMPORARY.
Select if (Var1="1").
FREQUENCIES
  VARIABLES=   InitialEigenvalues_Total
  /FORMAT NOTABLE
  /PERCENTILES= 2.5 97.5
  /STATISTICS=STDDEV MINIMUM MAXIMUM MEAN MEDIAN
  /HISTOGRAM NORMAL.
```

> Change highlighted OMS commands using OMS utility in SPSS to get different data from EFA

> This section actually performs the 5000 bootstrap analyses of EFA.

SPSS Syntax to align absolute values of factor loadings into columns for bootstrapping

```
String Factor (A2).
Compute factor= substr(var1, 1, 2).
execute.
compute F1=abs(@1).
compute F2=abs(@2).
compute F3=abs(@3).
execute.
*move Math loadings to F1.
do if (Factor = "Ma").
do if (abs(@3)>abs(@1)).
Compute F1=abs(@3) .
compute F3=abs(@1) .
end if.
end if.
execute.
do if (Factor = "Ma").
do if (F2>F1).
compute junk=F1.
Compute F1=f2.
compute f2=junk.
recode junk (lo thru hi=sysmis).
end if.
end if.
execute.
*move Eng loadings to F3.
do if (Factor = "En").
do if (abs(@3)<abs(@1)).
Compute F1=abs(@3) .
compute F3=abs(@1) .
end if.
end if.
execute.
do if (Factor = "En").
do if (F2>F3).
compute junk=F3.
Compute F3=f2.
compute f2=junk.
recode junk (lo thru hi=sysmis).
end if.
end if.
execute.
***move PA to F2.
do if (Factor = "Pa").
do if (abs(@3)>abs(@2)).
compute F2=abs(@3) .
compute F3=abs(@2) .
end if.
```

```
end if.
execute.
do if (Factor = "Pa").
do if (f1>f2).
compute junk=F1.
Compute F1=f2.
compute f2=junk.
recode junk (lo thru hi=sysmis).
end if.
end if.
execute.
```

6 DATA CLEANING AND EFA

If this is not the first work of mine you have come across, you might know that I have been a constant (perhaps tiresome) advocate of the argument that data are not ready to analyze until they are clean and missing data are dealt with. My second book was entirely about all the different –legitimate- things a researcher can do to improve the quality of their data and the results that come from analysis of those data.

Exploratory factor analysis is no exception, but there are different issues, and different priorities when dealing with EFA. In many inferential statistics, we can utilize tools like residuals to help us identify cases that are inappropriately influential. With EFA being an exploratory technique, we do not have those types of tools to work with. Nevertheless, in this chapter I will briefly review some data cleaning issues relevant to this analytic technique.

Two types of outliers in EFA: individual cases and variables

In factor analysis, there are actually two different types of outliers.[38] The first type of outlier is a case (or value) that does not belong. There are many reasons why cases become outliers: the case could be from a different population, could be the result of data recording or entry error, could have resulted from motivated mis-responding, or could represent a small subgroup with a different factor structure that has not been recognized. Whatever the reason, having these "illegitimate" cases in the data does not serve any useful purpose. Values that do not belong can arise from data entry errors (such as a 9 accidentally entered for a Likert-type item that only has values from 1-6). If the original data is available, you can check the data manually and fix the problem. If this is not possible, you could remove it and then use missing-data techniques

---

[38] In this context, I will use the term "outlier" to loosely describe a thing that does not belong. There are more technical definitions for different contexts, but you will have to indulge me here. There are also some in the scholarly community who insist that removing outliers (and indeed, data cleaning in general) is detrimental to the scientific process. I have published many articles, given numerous talks, and ultimately, wrote an entire book to empirically demonstrate why this position is wrong-headed.

(described below) to replace it with a reasonable estimate of what it might have been.

The second type of outlier in EFA is when a variable is an outlier. In this context, a variable can be considered an outlier when it loads on its own factor as a single-item factor. If an item does not load on any other factor, and no other items load on that factor, it is considered an outlier, and should be removed from the analysis (or the scale should be reworked to more fully represent that dimension of the latent construct if it is a legitimate facet of the construct.

The second type of outlier is easier to identify, obviously. There is a variable in your analysis that does not play well with others. Remove it and try the analysis again. The accidental value that is out of bounds is easy to identify by perusing frequency distributions for the variables to be analyzed. This should always be a first step for researchers.

The individual case that is an outlier is more tricky to identify in EFA. Visual examination of the data can help one identify odd patterns (like the participant who answers "3" to every item). Other patterns, like random responding or motivated mis-responding are more difficult to identify.

Response sets and unexpected patterns in the data[39]

Response sets can be damaging to factor analysis and to the quality of measurement in research. Much of the research we as scientists perform relies upon the goodwill of research participants (students, teachers, participants in organizational interventions, minimally-compensated volunteers, etc.) with little incentive to expend effort in providing data to researchers. If we are not careful, participants with lower motivation to perform at their maximum level may increase the error variance in our data, masking real effects of our research. In the context of this book, random and motivated mis-responding can have deleterious effects such as masking a clear factor structure or attenuating factor loadings and communalities.

Response sets (such as random responding) are strategies that individuals use (consciously or otherwise) when responding to educational or psychological tests or scales. These response sets range on a continuum from unbiased retrieval (where individuals use direct, unbiased recall of factual information in memory to answer questions) to generative strategies (where individuals create responses not based on factual recall due to inability or unwillingness to produce relevant information from memory; see Meier 1994, p. 43). Response sets have been discussed in the measurement and research methodology literature for over seventy years now (Cronbach, 1942; Goodfellow, 1940; Lorge, 1937), and some (e.g., Cronbach, 1950) argue that response sets are ubiquitous, found in almost every population on almost every type of test or assessment. In fact, early researchers identified response sets on assessments as diverse as the Strong Interest Inventory (Strong, 1927), tests of clerical aptitude, word meanings, temperament, and spelling, and judgments of proportion in color mixtures, seashore pitch, and pleasantness of stimuli, (see summary in Cronbach, 1950, Table 1).

[39] Parts of this section are adapted from Osborne, J.W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology, Vol 1, Article 220, pp. 1-7* doi: 10.3389/ fpsyg.2010.00220.

Researchers (myself included) are guilty of too often assuming respondents exclusively use unbiased retrieval strategies when responding to questionnaires or tests, despite considerable evidence for the frequent use of the less desirable and more problematic generative strategies (Meier, 1994; pp. 43-51).

Commonly discussed response sets

Examples of common response sets discussed in the literature include:

Random responding is a response set where individuals respond with little pattern or thought (Cronbach, 1950). This behavior, which completely negates the usefulness of responses, adds substantial error variance to analyses. Meier (1994) and others suggest this may be motivated by lack of preparation, reactivity to observation, lack of motivation to cooperate with the testing, disinterest, or fatigue (Berry et al., 1992; Wise, 2006). Random responding is a particular concern in this paper as it can mask the effects of interventions, biasing results toward null hypotheses, smaller effect sizes, and much larger confidence intervals than would be the case with valid data.

Malingering and dissimulation. Dissimulation refers to a response set where respondents falsify answers in an attempt to be seen in a more negative or more positive light than honest answers would provide. Malingering is a response set where individuals falsify and exaggerate answers to appear weaker or more medically or psychologically symptomatic than honest answers would indicate, often motivated by a goal of receiving services they would not otherwise be entitled to (e.g., attention deficit or learning disabilities evaluation; Kane (2008); see also Rogers, 1997) or avoiding an outcome they might otherwise receive (such as a harsher prison sentence; see e.g., Ray, 2009; Rogers, 1997). These response sets are more common on psychological scales where the goal of the question is readily apparent (e.g., "Do you have suicidal thoughts?";see also Kuncel & Borneman, 2007). Clearly, this response set has substantial costs to society when individuals dissimulate or malinger, but researchers should also be vigilant for these response sets, as motivated responding such as this can dramatically skew research results.

Social desirability is related to malingering and dissimulation in that it involves altering responses in systematic ways to achieve a desired goal—in this case, to conform to social norms or to "look good" to the examiner (see, e.g., Nunnally & Bernstein, 1994). Many scales in psychological research have attempted to account for this long-discussed response set (Crowne & Marlowe, 1964), yet it remains a real and troubling aspect of research in the social sciences that may not have a clear answer, but can have clear affects for important research (e.g., surveys of risky behavior, compliance in medical trials, etc.).

Other response styles such as acquiescence and criticality, are response patterns wherein individuals are more likely to agree with (acquiescence) or disagree with (criticality) questionnaire items in general, regardless of the nature of the item (e.g., Messick, 1991; Murphy & Davidshofer, 1988).

Response styles peculiar to educational testing are also discussed in the literature. While the response styles above can be present in educational data, other biases peculiar to tests of academic mastery (often multiple choice) include: (a) response bias for particular columns (e.g., A or D) on multiple choice type items, (b) bias for or against guessing when uncertain of the correct answer, and (c) rapid guessing (Bovaird, 2003), which is a form of random responding discussed above. As mentioned above, random responding (rapid guessing) is undesirable as it introduces substantial error into the data, which can suppress the ability for researchers to detect real differences between groups, change over time, and the effect(s) of interventions.

Summary. We rely upon quantitative research to inform and evaluate instructional innovations, often with high stakes and financial implications for society as a whole. Some interventions involve tremendous financial and time investment (e.g., instructional technology, community outreach agencies), and some might even be harmful if assessed validly, and therefore can be costly to individuals in terms of frustration, lost opportunities, or actual harm. Thus, it is important for researchers to gather the best available data on interventions to evaluate their efficacy. Yet research must rely upon the good faith and motivation of participants (students, teachers, administrators, parents, etc.) for which they may find neither enjoyment nor immediate benefit. This leaves us in a quandary of relying on research to make important decisions, yet often having flawed data. This highlights the importance of all data cleaning (including examining data for response bias) in order to draw the best possible inferences. This paper, and our example, focuses on educational research, but the lesson should generalize to all social sciences research (and beyond).

Is random responding truly random?

An important issue is whether we can be confident that what we call "random responding" truly is random, as opposed to some other factor affecting responses. In one study attempting to address this issue, Wise (2006) reported that answers identified as random responding on a four-choice multiple choice test (by virtue of inappropriately short response times on a computer based tests) were only correct 25.5% of the time, which is what one would expect for truly random responses in this situation. On the same test, responses not identified as random responding (i.e., having appropriately long response times) were correct 72.0% of the time.[40] Further, this issue does not appear to be rare or isolated behavior. In Wise's (2006) sample of university sophomores, 26% of students were identified as having engaged in random responding, and Berry et al. (1992) reported the incidence of randomly responding on the MMPI-2 to be 60% in college students, 32% in the general adult population, and 53% amongst applicants to a police training program. In this case, responses identified as random were more likely to be near the end of this lengthy assessment, indicating these

---

[40] Wise utilized computer based testing, allowing him to look at individual items rather than students' total test score. While computer-based testing can eliminate some aspects of random responding, such as choosing illegitimate answers, it does not eliminate random selection of items, or rapid guessing.

responses were likely random due to fatigue or lack of motivation.

In my study on this topic (Osborne & Blanchard, 2011), we found that about 40% of the 560 students involved in a study designed to assess the effects of an educational intervention were engaging in motivated mis-responding – in this case, probably random responding. They were identified by two different criteria, discussed below: Rasch outfit measures and performance on a random responding scale. To confirm the label, we demonstrated that random responders received substantially lower test scores than other students, and also showed much less change over time (before vs. after intervention) compared to other students. We went through other analyses to validate that those identified as random responders were indeed random responders that I will not go into now, as it is not central to the point of this chapter.
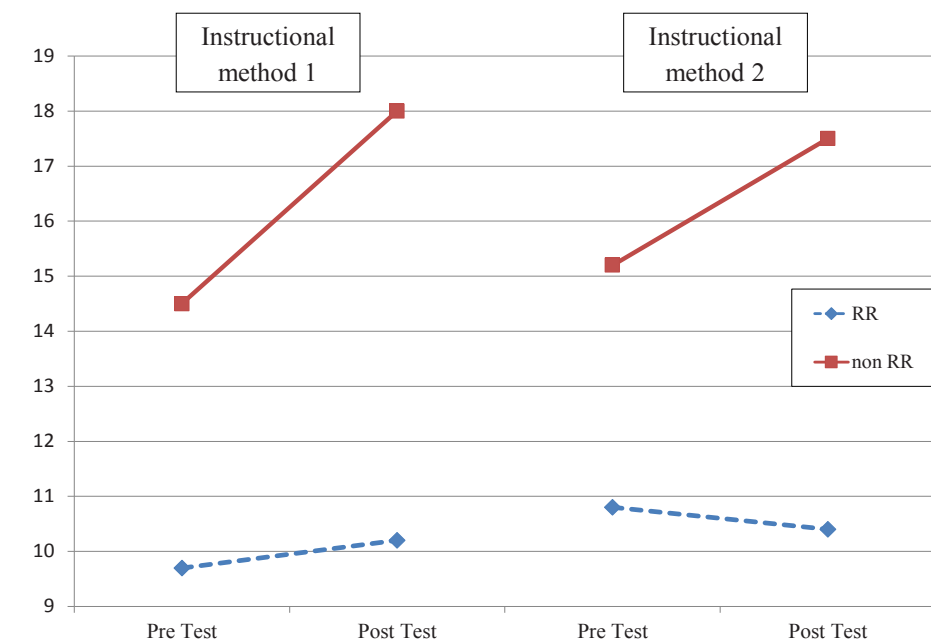
Figure 6.1: Random responders vs. other students' performance on an educational test before and after an educational intervention.

Detection of random responding

There is a well-developed literature on how to detect many different types of response sets that goes far beyond the scope of this paper to summarize. Examples include addition of particular types of items to detect social desirability, altering instructions to respondents in particular ways, creating equally desirable items worded positively and negatively, and for more methodologically sophisticated researchers, using IRT to explicitly estimate a guessing (random response) parameter. Meier (1994; see also Rogers, 1997) contains a succinct summary of some of the more common issues and recommendations around response set detection and avoidance. However, our example and focus for the rest of this paper will remain one of the most damaging

common response sets (from an inference perspective): random responding.

Creation of a simple random responding scale. For researchers not familiar with IRT methodology, it is still possible to be highly effective in detecting random responding on multiple choice educational tests (and often on psychological tests using likert-type response scales as well). In general, a simple random responding scale involves creating items in such a way that 100% or 0% of the respondent population should respond in a particular way, leaving responses that deviate from that expected response suspect. There are several ways to do this, depending on the type of scale in question. For a multiple-choice educational test, one method (most appropriate when students are using a separate answer sheet, such as a machine-scored answer sheet, used in this study, and described below) is to have one or more choices that are illegitimate responses.[41]

A variation of this is to have questions scattered throughout the test that 100% of respondents should answer in a particular way if they are reading the questions (Beach, 1989). These can be content that should not be missed (e.g., 2+ 2= ___), behavioral/attitudinal questions (e.g., I weave the fabric for all my clothes), nonsense items (e.g., there are 30 days in February) or targeted multiple choice test items, such as:

How do you spell 'forensics'?
(a) fornsis,
(b) forensics,
(c) phorensicks,
(d) forensix).

Item response theory. One application of item response theory has implications for identifying random responders using item-response theory to create person-fit indices (Meijer, 2003). The idea behind this approach is to quantitatively group individuals by their pattern of responding, and then use these groupings to identify individuals who deviate from an expected pattern of responding. This could lead to inference of groups using particular response sets, such as random responding. Also, it is possible to estimate a "guessing parameter" and then account for it in analyses, as mentioned above.

A thorough discussion of this approach is beyond the scope of this article, and interested readers should consult references such as Edelen and Reeve (2007; see also Hambleton, Swaminathan, & Rogers, 1991; Wilson, 2005). However, IRT does have some drawbacks for many researchers, in that it generally requires large (e.g., N ≥ 500) samples, significant training and resources, and finally, while it does identify individuals who do not fit with the general response pattern, it does not necessarily show what the response set, if any, is. Thus, although useful in many instances, we cannot use it for our study.

---

[41] One option, used in this particular data set included having twenty questions with four choices: A-D, with other questions scattered throughout the test, and particularly near the end, with items that contain only three (A-C) or two (A-B) legitimate answers. Students or respondents choosing illegitimate answers one or more times can be assumed to be randomly responding, as our results show.

Rasch measurement approaches. Rasch measurement models are another class of modern measurement tools with applications to identifying response sets. Briefly, Rasch analyses produce two fit statistics of particular interest to this application: infit and outfit, both of which measure sum of squared standardized residuals for individuals.[42] Large infit statistics can reflect unexpected patterns of observations by individual (usually interpreted as items mis-performing for the individuals being assessed), while large outfit mean squares can reflect unexpected observations by persons on items (may be the result of haphazard or random responding). Thus, large outfit mean squares can indicate an issue that deserves exploration, including haphazard or random responding.

Again, the challenge is interpreting the cause (response set or missing knowledge, for example, in an educational test) of the substantial outfit values. We will use this application of Rasch as a check on the validity of our measure of random responding below. Again, a thorough discussion of this approach is beyond the scope of this article but interested readers can explore Bond and Fox (2001) and/or Smith and Smith (2004).

Summary. No matter the method, we assert that it is imperative for educational researchers to include mechanisms for identifying random responding in their research, as random responding from research participants is a threat to the validity of educational research results. Best practices in response bias detection is worthy of more research and discussion, given the implications for the quality of the field of educational research. In order to stimulate discussion and to encourage researchers to examine their data for this issue, we share an example from our own research demonstrating how a small number of individuals engaging in random responding can mask the effects of educational interventions, decreasing researchers' ability to detect real effects of an educational intervention.

An example of the effect of random or constant responding

In this section we will review two types of problematic responding: random responding and constant responding. Random responding is randomly entering numbers within a given range. Humans tend not to engage in true random responding, however. So the other extreme is for individuals to respond with a constant number across all questions (such as "3"). For each of the examples below, 100 of the 300 cases in the small Marsh SDQ data set (the same one we bootstrapped in Chapter 5), were either replaced with randomly generated cases with a uniform random distribution from 1-6 (integers only), or with a "3" to represent a 33% random responding or constant responding rate.

As you can see in Table 6.1, the effect is noticeable. When 33% of a sample is engaging in random responding on an instrument that has very strong and clear factor structure, the initial *and* extracted eigenvalues are attenuated, as is common in most

---

[42] infit is an abbreviation for "information weighted mean square goodness of fit statistic" and outfit is an abbreviation for "outlier sensitive mean square residual goodness of fit statistic," (Smith and Smith, 2004, p. 13)

statistical procedures when random error is introduced into the data. The extracted variance in the random condition is reduced 26.83% compared to the original data (44.17% vs. 60.50%).

Table 6.1a

Effects of random or constant responding on Marsh SDQ data

	Original		33% random		33% constant	
Factor:	Initial	ML Extraction	Initial	ML Extraction	Initial	ML Extraction
1	4.082	3.399	3.891	3.296	5.888	5.486
2	2.555	2.446	1.800	1.361	2.041	1.843
3	2.208	1.874	1.637	1.086	1.728	1.435
4	.908		.890		.786	
5	.518		.797		.635	
6	.487		.671		.436	
% var	68.83%	60.50%	56.37%	44.17%	74.28%	67.41%

Table 6.1b

Effects of random or constant responding on factor loadings

	Sample (N=300) Factor Loadings			Random responding sample			Constant responding sample		
	1	2	3	1	2	3	1	2	3
Math1	.916	-.052	-.006	.815	-.038	-.015	.967	-.109	-.005
Math2	.875	.001	-.034	.758	.032	.003	.901	.008	-.009
Math3	.875	.038	-.010	.681	-.002	.063	.851	.070	.022
Math4	-.508	-.030	-.083	-.337	-.247	.023	-.525	-.106	.002
Par1	.055	.753	-.109	.041	.721	-.049	.069	.815	.046
Par2	.043	-.626	-.003	.007	-.600	-.010	.055	-.745	.022
Par3	-.034	.873	.049	-.031	.736	-.002	-.033	.897	.060
Par4	.023	-.592	-.113	.026	-.574	-.151	.013	-.583	-.058
Par5	.064	.742	-.023	.090	.578	-.095	.088	.816	-.123
Eng1	.033	.035	.770	.099	.034	.541	.081	.061	.805
Eng2	.025	-.082	.859	-.149	.016	.740	-.153	.007	.931
Eng3	.052	.000	.855	-.038	.079	.688	-.022	.041	.870
Eng4	.088	-.073	-.675	-.255	.144	-.446	-.180	.114	-.691

Note: ML extraction with Promax rotation.

In the constant condition, the initial eigenvalue is increased dramatically, as this pattern of responding makes the data look much closer to a single strong factor solution. The other eigenvalues are attenuated, yet the overall variance accounted for is inappropriately increased due to the first factor.

Random data can also lead to potential confusion about the factor structure. For example, according to MAP Criteria, the data with 33% random recommends extraction of two factors according to the revised (2000) map criteria (although the original 1976 criteria still recommend extraction of three factors). The constant responding did not introduce this issue into the MAP analysis, at least with this low a percentage of cases exhibiting the response set.

As you can see in Table 6.1b, the factor loadings for the random sample are attenuated, sometimes markedly (e.g., English items). In the constant responding data, many of the factor loadings are inflated. Although the basic factor pattern is present in all three data sets despite these challenges, in a more marginal data set this might make a difference between a clear factor structure and an analysis that does not produce the appropriate factor structure. Replication might also be an issue, particularly if the portion of respondents engaging in random or constant responding changes markedly.

Data cleaning

Unless there is an odd pattern to outliers, failure to check for data quality issues would lead to similar effects as the random responding data, depending on how egregious the outliers are. In general, outliers introduce error variance into the analysis, as random responding does. Thus, we should see similar results.

Where one might see a difference is in the common case of government data sets, where missing data values are often something like 98 or 99. In the case like this where all variables should range from 1-6, values of 98 or 99 can cause a large amount of error variance to be introduced into the analysis.

Missing data

Missing data is an issue in exploratory factor analysis as EFA will only analyze complete cases, and thus any case with missing data will be deleted. This can reduce sample size, causing estimates to be more volatile. If missingness is random, then your estimates should be unbiased. However, it is unusual for missing data to be completely at random. Thus, it is likely that missing data is causing bias in the results *in addition to* reducing sample size—unless you deal with the missing data in some appropriate manner.

What Is Missing or Incomplete Data? If any data on any variable from any participant is not present, the researcher is dealing with missing or incomplete data. In many types of research, it is the case that there can be *legitimate missing data*. This can come in many forms, for many reasons. Most commonly, legitimate missing data is an absence of data when it is appropriate for there to be an absence. Imagine you are filling out a survey that asks you whether you are married, and if so, how long you have been married. If you say you are not married, it is legitimate for you to skip the follow-up question on how long you have been married. If a survey asks you whether you voted in the last election, and if so, what party the candidate was from, it is legitimate to skip the second part if you did not vote in the last election.

Legitimately missing data can be dealt with in different ways. One common way of dealing with this sort of data could be using analyses that do not require (or can deal effectively with) incomplete data. These include things like hierarchical linear modeling

(HLM; Raudenbush & Bryk, 2002) or survival analysis. Another common way of dealing with this sort of legitimate missing data is adjusting the denominator. Again taking the example of the marriage survey, we could eliminate non-married individuals from the particular analysis looking at length of marriage, but would leave non-married respondents in the analysis when looking at issues relating to being married versus not being married. Thus, instead of asking a slightly silly question of the data—"How long, on average, do all people, even unmarried people, stay married?"—we can ask two more refined questions: "What are the predictors of whether someone is currently married?" and "Of those who are currently married, how long on average have they been married?" In this case, it makes no sense to include non-married individuals in the data on how long someone has been married.

Illegitimately missing data is also common in all types of research. Sensors fail or become mis-calibrated, leaving researchers without data until that sensor is replaced or recalibrated. Research participants choose to skip questions on surveys that the researchers expect everyone to answer. Participants drop out of studies before they are complete. Missing data also, somewhat ironically, can be caused by data cleaning (if you delete outlying values).

Few authors seem to explicitly deal with the issue of missing data, despite its obvious potential to substantially skew the results (Cole, 2008). For example, in a recent survey my students and I performed of highly regarded journals from the American Psychological Association, we found that more than one-third (38.89%) of authors discussed the issue of missing data in their articles. Do those 61% who fail to report anything relating to missing data have complete data (rare in the social sciences, but possible for some authors), do they have complete data because they removed all subjects with any missing data (undesirable, and potentially biasing the results, as we discuss below), did they deal effectively with the missing data and fail to report it (less likely, but possible), or did they allow the statistical software to treat the missing data via whatever the default method is, which most often leads to deletion of subjects with missing data? If our survey is representative of researchers across the sciences, we have cause for concern. Our survey found that of those researchers who did report something to do with missing data, most reported having used the classic methods of listwise deletion (complete case analysis) or mean substitution, neither of which are best practices (Schafer & Graham, 2002). In only a few cases did researchers report doing anything constructive with the missing data, such as estimation or imputation.

Regression and multiple imputation have emerged as two more progressive methods of dealing with missing data, particularly in cases like factor analysis where there are other closely correlated variables with valid data. Regression imputation (also referred to as simple imputation) creates a regression equation to predict missing values based on variables with valid data. This has been shown to be superior to mean substitution or complete case analysis, particularly when data are not missing completely at random.

Multiple imputation uses a variety of advanced techniques—e.g., EM/maximum likelihood estimation, propensity score estimation, or Markov Chain Monte Carlo (MCMC) simulation—to estimate missing values, creating multiple versions of the same data set (sort of a statistician's view of the classic science fiction scenario of alternate realities or parallel universes) that explore the scope and effect of the missing

data. These parallel data sets can then be analyzed via standard methods and results combined to produce estimates and confidence intervals that are often more robust than simple (especially relatively weak) imputation or previously mentioned methods of dealing with missing values (Schafer, 1997, 1999).

The effects of nonrandom missingness. To simulate some non-random missingness, I recoded "1" or "2" to system missing values for the first three English items:

Eng1: I learn things quickly in English classes
Eng2: English is one of my best subjects
Eng3: I get good marks in English

This created a biased sample eliminating those small numbers of students who answered the most pessimistically on the English items (67 out of 300 cases).

Table 6.2a
Effects of nonrandom missing data (N=67) on Marsh SDQ data

Factor:	Original		Nonrandom missing	
	Initial	ML Extraction	Initial	ML Extraction
1	4.082	3.399	4.242	3.891
2	2.555	2.446	2.582	2.179
3	2.208	1.874	1.904	1.690
4	.908		1.267	.953
5	.518		.561	
6	.487		.466	
% variance	68.83%	60.50%	76.88%	67.02%

As you can see in Table 6.2a, with some cases deleted from the analysis in a non-random fashion we suddenly have a different factor structure. We now have four factors with an eigenvalue greater than 1.0, and MAP criteria confirm that in this data set, four is the recommended number of factors that should be extracted. Because of the extra factor extracted, the variance accounted for is somewhat inflated.

As you can see in Table 6.2b, the new factor structure is mostly intact for math and English, but the parent factor is now split into two factors. The fourth factor seems to represent some sort of general negativity. It is not uncommon in EFA for negatively worded items to load on a separate factor from positively worded items on the same latent construct, but of course this is not ideal nor is it consistent with the theoretical framework.

Table 6.2b
Effects of random or constant responding on factor loadings

	Original Marsh SDQ sample			Non-random missing			
	1	2	3	1	2	3	4
Math1	.916	-.052	-.006	.022	.881	.021	.080
Math2	.875	.001	-.034	.078	.825	.058	.123
Math3	.875	.038	-.010	-.015	.882	-.007	-.084
Math4	-.508	-.030	-.083	.140	-.509	.098	.438
Par1	.055	.753	-.109	.797	.028	.052	.011
Par2	.043	-.626	-.003	-.359	.074	.140	.667
Par3	-.034	.873	.049	.927	-.059	.102	-.019
Par4	.023	-.592	-.113	-.346	-.061	.064	.536
Par5	.064	.742	-.023	.747	.064	-.028	-.030
Eng1	.033	.035	.770	.096	-.051	.728	-.048
Eng2	.025	-.082	.859	-.006	-.021	.804	.069
Eng3	.052	.000	.855	.048	.127	.813	-.077
Eng4	.088	-.073	-.675	.232	.091	-.324	.752

Note: ML extraction with Promax rotation.

Summary. The tradition within quantitative methods (and indeed, the default in statistical computing packages) is to delete cases with missing data. However, this is almost never the ideal solution, even if the assumption that the data are missing completely at random is met (which it almost never is). When data are missing not at random, you may be introducing serious bias into your results by ignoring this issue. Simulations presented in my chapter on missingness (Osborne, 2013, Chapter 6) shows that modern methods of dealing with missing data (simple or multiple imputation, for example) can effectively ameliorate the harmful effects of nonrandom missingness- and as a bonus, keep all those hard-earned data points in your analysis.

Chapter 6 Conclusions

Data quality is a continual issue in almost all sciences. In EFA, as in most other quantitative analyses, data quality issues can bias or completely derail your analyses. Thus, I encourage you to closely examine all data prior to analysis for extreme cases, random or other types of motivated mis-responding, and to deal with missing data effectively.

7 ARE FACTOR SCORES A GOOD IDEA?

Factor scores seemed cutting-edge back in the 1980s when I was beginning to take graduate statistics courses. The concept and practice of computing factor scores extends back to the 1920s, although early on it was considered a much less important aspect of factor analysis than the determination of the actual factors themselves. In the early decades of the 20th century there also was extended arguments about whether it was even proper to compute factor scores, due to something called "indeterminancy"- essentially meaning that there are more unknowns than equations being estimated in EFA (for an excellent overview of the issue, see Grice, 2001) and the same individuals could be ranked multiple ways, leaving their relative ranking indeterminant. I will not delve into this issue further because I do not believe that factor scores are useful for us to be calculating, as you will see as the chapter plays out.

In essence, what a factor score is an approximation of what an individual might score on a latent construct/factor. There are several methods for calculating factor scores. The most common is to simply sum or average scores from a scale. Most researchers do this without considering the implicit assumption that action entails- that all items are equally weighted. In other words, when you sum or average items, you are explicitly asserting that all items contribute equally to the construct, and thus, they are appropriate for averaging or summing. However, decades ago, researchers using factor analysis started thinking about this assumption, and realizing that EFA explicitly shows that this assumption is often not warranted. Some items are more indicative of a construct than others. Thus, researchers started weighting items according to the results of EFA analyses—and in fact, modern statistical software packages often include this as an option when performing an EFA.

Attempting to improve measurement, researchers could weight each variable differently as a function of the strength of its loading and then sum to approximate what might be the true score of each individual on the construct being examined.[43] In

---

[43] refined techniques are actually a bit more complex, taking into account more information than just factor loading—for example, the correlation between factors. However, as the point of this chapter is to discourage the use of factor scores, I will refrain from providing more detail.

structural equation modeling today, we can also use those capabilities to save and analyze scores on latent constructs (or just analyze the latent constructs as variables). It makes sense and was a natural progression over a period of many decades of research.

However, we now have a different assumption: that the factor loadings (correlations between an item and the factor) are stable and generalizable. This issue is similar to that of researchers using multiple regression to predict outcomes for individuals (for papers on prediction in regression, see Osborne, 2000, 2008a). Specifically, the issue is that these procedures *overfit* the data. Most samples contain idiosyncratic aspects that most quantitative analyses will take advantage of to fit the model, despite those sample characteristics being non-reproducible (Thompson, 2004, p. 70). Because EFA suffers from the same overfitting, I recommended in earlier chapters that we focus on replication and evaluate the anticipated precision or variability across samples to evaluate the goodness of EFA results.

This is going to be a short chapter for the following reasons:

1. I think I have clearly staked a position that EFA should be used as an exploratory technique only, and as prelude to follow-up with confirmatory methods or modern measurement analyses like Rasch or IRT methods. Thus, from a philosophical point of view, I find factor scores problematic in that EFA was not designed to produce highly refined estimates of latent variables for use in subsequent analyses.
2. The solutions that we often receive from EFA are highly unstable across samples, and thus factor scores would be highly unstable. This is not a good situation for scientific inquiry
3. There are excellent resources for researchers to refer to about the technical details of factor scores, various options, and potential issues (e.g., DiStefano, Zhu, & Mindrila, 2009). However, in light of points #1 and 2, above, this information is not a good use of our time.

Pedagogical Example: Engineering data

We will start with the same first example from chapter 2: the engineering data. I have copied the items and information below so you do not have to flip back and forth between this chapter and chapter 2.

Table 7.1
Factor loadings matrix from engineering data

Variable:	Rotated pattern coefficients	
	1	2
EngProbSolv1	.859	-.016
EngProbSolv2	.841	-.071
EngProbSolv3	.879	-.008
EngProbSolv4	.909	-.025
EngProbSolv5	.886	.021
EngProbSolv6	.869	.020
EngProbSolv7	.868	.033
EngProbSolv8	.790	.072
INTERESTeng1	.042	.801
INTERESTeng2	-.023	.921
INTERESTeng3	-.014	.922
INTERESTeng4	-.001	.904
INTERESTeng5	-.007	.897
INTERESTeng6	.009	.864

Proper vs. improper factor scores

One issue of terminology will be the issue of whether you compute factor scores using all the variables (a proper factor score) or just the variables that compose a particular factor. Proper factor scores take into account *all variables* in the analysis, not just the ones that are considered to be part of a factor. So, referring to Table 7.1, a proper factor score for the first factor would include both the engineering problem solving variables and the identification with engineering variables. Of course, those second loadings are relatively small, and as such, would contribute very little to the analysis. An improper factor score would include only those variables considered to be part of the factor. In fact, authors have shown that these two types of factor scores are generally highly correlated, and that improper factor scores are a bit more replicable and robust. And both, in my opinion, are equally flawed.

How unstable are factor scores?

It is arguable whether factor scores are actually an improvement over equally-weighted composites (averaging, summing). On one hand, factor scores do account for the fact that some items contribute much more strongly to constructs than others (this is also part of the fundamental approach of Rasch measurement). On the other hand, if we care at all about replicability, and if factor loadings are not likely to be replicable, we could be causing more harm than good by using this approach. Of course, we have already established that smaller samples have more variability than large samples, so let us examine a few random samples at the 10:1 (participant: item) ratio (N= 140). [44]

Table 7.2
Variability in engineering data pattern coefficients across four random samples

Variable:	Sample #			
	1	2	3[1]	4
EngProbSolv1	.83	.86	.79	.89
EngProbSolv2	.86	.73	.59	.93
EngProbSolv3	.90	.93	.87	.74
EngProbSolv4	.92	.91	.77	.96
EngProbSolv5	.89	.80	.77	.86
EngProbSolv6	.76	.55	.99	.86
EngProbSolv7	.80	.51	.98	.77
EngProbSolv8	.61	.75	.72	.91
Identification with engineering scale				
INTERESTeng1	.90	.87	1.02	.81
INTERESTeng2	.99	.93	.95	.97
INTERESTeng3	.78	.85	1.01	.83
INTERESTeng4	.98	.97	.81	.91
INTERESTeng5	.83	.93	.94	.95
INTERESTeng6	.98	.92	.87	.81

1. In the third sample, we observe some values over 1.0. This is usually a sign of trouble, but SPSS would calculate factor scores anyway.

The pattern matrix for both the engineering problem solving and interest in engineering scales are presented in Table 7.2. Examining these four small random samples, you should see enough variability to encourage you to be skeptical the concept of factor scores. For example, the sixth item in engineering problem solving ranges from 0.55 to 0.99 across the four samples. Several other items have similar ranges. Remember that this is a scale that had very strong psychometric properties.

If you find yourself unconvinced, review the previous chapter on bootstrap resampling and review how volatile some of the other scales can be. This scale seems to be solid and stable even at low sample sizes. However, it is probably much more stable than most. You can also refer to Grice's (2001) excellent work on the topic for more information.

What are modern alternatives?

Structural Equation modeling. As I have already mentioned, structural equation modeling explicitly models latent variables, while factor scores tend to estimate what individuals *might* score on a factor. If the goal is to attempt to understand how latent variables (constructs) relate to each other, we have been able to directly model this for

---

[44] of course this sample size is entirely too small, but 10:1 ratio is about average for most EFAs reported in the literature.

the better part of a quarter century. I would encourage you to do just that. However, SEM has many of the same drawbacks in terms of replication that we just discussed in relation to multiple regression and EFA. Most notably, it will tend to overfit the model to the sample, so if the sample is quirky, small, or contains biases or error, the solution will not be as generalizable as we would hope. So SEM is not a panacea. You must have an unbiased, large sample in order to hope for replicability—and then you should test whether the results replicate or not.

Rasch (or IRT) modeling. Rasch measurement is similar to item response theory in that it seeks to understand how patterns of responses across items of different "difficulty" can help estimate a person score on a construct. I am more familiar with Rasch than IRT, and although they have similarities, scholars in both groups will tell you they are different in important ways. I leave it to you to pursue, as there are great books on both. For example, I find Bond and Fox (2006) an accessible and helpful book on Rasch measurement.

Chapter 7 Summary

I attempted to keep this chapter short as I think in the 21st century pantheon of quantitative methods, factor scores really don't have a legitimate place. Given what we know about the volatility of EFA analyses, even when the scale is traditionally strong with an unusually clear factor structure, and the conceptual and mathematical issues with factor scores (e.g., controversial mathematical issues like indeterminancy dating back to the early 20th century), I think we should close this chapter by affirming that we will use modern methods of modeling latent variables (e.g., SEM, Rasch, IRT) instead.

8 HIGHER ORDER FACTORS

Whenever factors are correlated, there is, naturally, a question as to whether there truly are several independent factors or whether there is a single "higher-order" factor.[45] This has been a point of discussion for many decades, and is often conceptually and theoretically important. For example, is self-concept a single thing, or several separate things? Is depression a single construct composed of several sub-constructs, or is it really not a coherent construct?

Scholars writing in this area since the early 20th century have argued that when initial factor analyses (we can refer to these as "first order" factors as then come from the first level of analysis) produce correlated factors, researchers should explore whether there are second- or higher-order factors in order to more fully explicate the model (e.g., Gorusch, 1983; Thompson, 2004).

There are at least two issues with higher-order factors that we need to address. First is how to perform the analysis and interpret the results. The second issue is more conceptual: if initial EFA produces abstractions (unobservable variables called factors) that might or might not be precise representations of population dynamics, in higher-order factor analysis we then propose to analyze these imperfect abstractions to create possibly more imperfect higher-order abstractions. This makes me a bit uneasy, particularly in an exploratory framework. Given how volatile and unpredictable the results of EFA can be, it seems that taking those results and analyzing them again doubles (or raises to a power) the risk of going awry and far afield of the true character of the population dynamics.

Since almost all factors are correlated in the population, the assertion that this analysis *needs* to take place under these conditions should be regarded carefully. Researchers must decide: (a) whether higher-order *exploratory* analyses are desirable, and (b) how strong a correlation warrants this extra exploration. If factors are correlated around $r = 0.25$ (about 6.25% overlap), is that enough of a correlation to justify higher-order analysis? What about $r = 0.40$, which equates to only 16% overlap? We do not

---

[45] Actually, my first question is usually whether the initial EFA got it right in asserting there were several "first-order" factors, or whether the analysis should have concluded there were fewer-or one – factors.

have a good sense of what a second-order factor really means, in my opinion.

Did the initial solution get it right?

In my mind, the primary issue to be decided *prior to a higher-order factor analysis* is whether the initial factor structure is appropriate or not. If we extract five factors, and then decide there is a single second-order factor (or even two second-order factors), the first question I would ask is whether the original solution was correct, or whether the correct first-order structure should have been one (or two) factors rather than five.[46]

In the seven or eight decades since this discussion began in earnest, many things have changed in quantitative methods. One is the easy access to confirmatory factor analysis techniques. Although we have not discussed confirmatory techniques, they are methods for directly testing hypotheses such as whether a particular data set is best characterized as one, two, or five factors. Thus, before launching into higher-order factor analysis, I would evaluate (and replicate) whether the initial solution was correct. I suspect that in many cases, the initial solution was indefensible or sub-optimal, and that the "higher-order factors" are really just the more parsimonious version of what should have been extracted initially. I would only explore higher-order factors after the initial factor structure has been thoroughly vetted through CFA as the most parsimonious and desirable. Of course, once in CFA, higher-order factors can be modeled and tested in that confirmatory framework![47]

If you want to explore this aspect of your data, in the spirit of intrepid exploration we can briefly cover some of the mechanics of the process.

Mechanics of performing second-order factor analysis in SPSS

In general, second-order factor analysis consists of analysis of correlation matrices. If you are performing principal components analysis, you can save component scores and examine them as variables in a second-order analysis, as only in PCA will the component correlations and component scores match exactly (Thompson, 2004, p. 73).[48] When using common factor analysis, we must analyze the correlation matrix.

To illustrate this methodology, I will use the engineering data we started exploring from Chapter 2, adding a third subscale that asked eight questions about feelings of belongingness in engineering. Using PAF extraction and Promax rotation (as before), there were 372 cases with valid data on all variables. The syntax to perform the higher-order analysis will be available through the book web site, and is annotated in Appendix A at the end of this chapter.

---

[46] This sort of question can spawn endless debate amongst scholars, and the literature is replete with examples of this type of debate. In my opinion, these (often vitriolic) debates fester because of the exploratory, volatile, and non-replicable nature of these analyses. If authors would quickly replicate or (ideally) move to confirmatory analyses, these debates are less apt to erupt as there are clear ways to test competing hypotheses.

[47] I tend to recommend Barbara Byrne's excellent reference on structural equation modeling (Byrne, 2010) for readers interested in CFA/SEM and higher-order factor analysis in a confirmatory context.

[48] This is not, in my mind, a reason to decide to use PCA, by the way.

First-order analysis of engineering data. The first eigenvalue was 9.58, the second was 3.73, the third was 1.84. This might indicate that there is a single factor, or based on theory, there might be three factors with a single, second-order factor. CFA would help test the first question. For the sake of expediency, let us assume three factor structure due to the strong theoretical basis.

The pattern and structure coefficients, presented in Table 8.1, show both a strong factor structure and some indication that there might be a higher order factor (or there should be a single factor). The structure coefficients show relatively strong loadings across factors, and the factors are moderately correlated (r range from 0.36 to 0.52).

Table 8.1
First-order factor loadings from engineering data

	Pattern Coefficients			Structure Coefficients		
	1	**2**	**3**	**1**	**2**	**3**
BELONGeng1	.283	.019	.504	.554	.368	.662
BELONGeng2	-.055	.130	.479	.242	.345	.514
BELONGeng3	-.042	-.088	.848	.369	.313	.783
BELONGeng4	-.087	-.025	.835	.340	.353	.778
BELONGeng5	.020	.094	.278	.198	.237	.334
BELONGeng6	.238	-.035	.626	.553	.357	.733
BELONGeng7	-.243	.392	.363	.087	.483	.428
BELONGeng8	.162	.100	.512	.465	.409	.645
EngProbSolv1	.856	-.026	.010	.852	.286	.444
EngProbSolv2	.833	-.075	.012	.813	.230	.411
EngProbSolv3	.880	-.018	.009	.879	.302	.460
EngProbSolv4	.929	-.012	-.046	.900	.298	.432
EngProbSolv5	.891	.021	-.011	.893	.335	.464
EngProbSolv6	.886	.033	-.041	.876	.331	.438
EngProbSolv7	.863	.039	-.005	.874	.345	.464
EngProbSolv8	.779	.052	.041	.820	.352	.474
INTERESTeng1	.038	.778	.053	.344	.817	.453
INTERESTeng2	.010	.931	-.053	.316	.909	.409
INTERESTeng3	.020	.929	-.041	.331	.916	.424
INTERESTeng4	.035	.922	-.059	.335	.906	.411
INTERESTeng5	-.014	.859	.068	.329	.887	.482
INTERESTeng6	.011	.850	.025	.329	.867	.448

Second order factor analysis. The correlation matrix was analyzed via EFA to determine whether it might be reasonable to assume a second-order factor that incorporates all three first-order factors (keep in mind my concerns over performing this analysis at all…). I used PAF extraction and would have used Promax rotation (authors such as Thompson recommend using oblique rotations for higher-order EFA) but as I asked for a single factor to be extracted, there was no rotation.

Table 8.2
Results of second-order factor analysis

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Factor	**Total**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	1.917	63.898	63.898	1.430	47.665	47.665
2	.643	21.438	85.336			
3	.440	14.664	100.000			

Factor Matrix

	Factor
	1
factor_1	.619
factor_2	.581
factor_3	.842

The results (presented in Table 8.2) support an argument that there is one single second-order factor, and that all three scales load moderately to strongly on it (ranging from 0.58 to 0.84). This latent variable seems to be more interest than problem-solving and belongingness, but you would have to ask my colleague who created the scale what it all means.

Replication example of second-order factor

While intriguing, and perhaps conceptually sensible, these results are not definitive, and the replicability of second-order factors are not well-studied. It is also possible (or likely) that this scale should have been a single-factor scale from the beginning. As mentioned previously, the best way to test this competing hypothesis is to perform confirmatory factor analysis comparing model fit between the three-factor model, the three-factor model with higher order factor, and a one-factor model, using a different, large sample. Then and only then would we be able to draw conclusions about the *best* fitting model for these data. For now, let us take a simple example of replication and see how well these results might replicate. Unfortunately this original sample is relatively small, so I randomly selected two samples of N=150 each (with replacement) from the original sample, and performed the same analysis twice to explore whether in this particular case, replication of second-order factors is a reasonable expectation. There are endless caveats to what I am about to do, including the lousy sample size. Let's explore anyway!

As you can see in Table 8.3a, the initial communalities were not terribly close, and not terribly far off. Remembering that these represent percent variance accounted for, factor 2, for example, ranges from about 30% accounted for to about 15%-- a wide margin. Likewise, the extracted communalities varied a bit. The extracted eigenvalues

were 1.36, 1.43, and 1.18 (representing 45.43%, 47.67%, and 39.43% variance accounted for, respectively). Table 8.3b shows the factor matrix from the two analyses. As you can see, they are not terribly different, but not terribly similar.

Table 8.3a
Communalities from second-order factor analysis

	Sample 1		Sample 2		Sample 3	
	Initial	**Extract**	**Initial**	**Extract**	**Initial**	**Extract**
factor_1	.220	.317	.286	.383	.200	.284
factor_2	.299	.456	.255	.337	.158	.222
factor_3	.335	.591	.378	.709	.276	.677

Table 8.3b
Factor loadings from second-order factor analysis

	Sample 1	Sample 2	Sample 3
factor_1	.563	.719	.533
factor_2	.675	.566	.471
factor_3	.769	.866	.823

Of note in the factor loadings table is the fact that these would be what we use to construct the meaning of the higher-order factor. Between the first two samples, the order of importance of the three factors changes. In the third sample, the third factor is even more strongly dominant and the second factor is even less important. This might be nitpicky, because I don't think EFA applied for this purpose is a best practice. If you are going to use this, I hope you replicate or bootstrap to give the reader an idea of how robust and/or precise your estimates are.

Chapter 8 Summary

Higher order factors are not commonly discussed in the literature, but they are present, and it is likely that you will run across discussion of this topic at some point. I included this chapter because I feel that it is not discussed enough in research methods, and too much in the content literature.

Higher order factors are tricky. They are not difficult to evaluate once you figure out how to perform the analyses (syntax will be provided on the book web site). The trickier issue is deciding whether there truly is a higher-order factor or whether the original analysis should have specified a single factor structure. I remain convinced that these exploratory techniques are fun, but relatively low in value without confirmatory techniques to evaluate the competing hypotheses in a more rigorous manner, and replication in independent samples. Also, if you are committed to performing this type of analysis within an EFA framework, Thompson (2004) has described in detail a more modern methodology for presenting this information.

Chapter 8 exercises

1. Download the engineering data along with the syntax for higher-order factor analysis and replicate the results presented above.
2. Examine the Marsh SDQ data and evaluate whether it appears there is a higher-order factor of "self concept" involved in those data.
3. Using your own or another data set, perform primary and secondary factor analyses.

Appendix 8A: Syntax for performing higher-order EFA

Below I have pasted the syntax I used to perform the analyses contained in the chapter. I will also have it available for download on the book web site. Note that this is closely adapted from syntax provided by IBM/SPSS on their web site (http://www-01.ibm.com/support/docview.wss?uid=swg21479182).

I also have highlighted the sections that you will need to examine or edit to perform the analyses yourself. If you keep the data and variables the same, you only need to edit the file locations. If you use this for another data set, you need to change the variable names, and perhaps also extraction and rotation details. You also, as described below, have to manually edit the extracted correlation matrix before saving in order for it to be suitable for SPSS to analyze. This is simple if you follow the directions, again provided kindly by SPSS on that web site above.

```
*This section sets up the extraction of correlation table, and performs the analysis.
*It will produce a new data set (correlation matrix) in another SPSS data file window.
OMS
/SELECT TABLES
/IF COMMANDS = ["Factor Analysis"]
SUBTYPES = ["Factor Correlation Matrix"]
/DESTINATION FORMAT = SAV
OUTFILE = "C:\temp\Factor_correlation_matrix.sav".
FACTOR
/VARIABLES  BELONGeng1 BELONGeng2 BELONGeng3 BELONGeng4
BELONGeng5 BELONGeng6 BELONGeng7 BELONGeng8 EngProbSolv1
EngProbSolv2 EngProbSolv3 EngProbSolv4 EngProbSolv5
EngProbSolv6 EngProbSolv7 EngProbSolv8 INTERESTeng1
INTERESTeng2 INTERESTeng3 INTERESTeng4 INTERESTeng5
INTERESTeng6
/MISSING LISTWISE
/PRINT INITIAL EXTRACTION ROTATION
/CRITERIA FACTORS(3) ITERATE(50)
/EXTRACTION PAF
/CRITERIA ITERATE(25)
/ROTATION PROMAX(4)
/METHOD=CORRELATION .
OMSEND.
GET FILE='C:\temp\Factor_correlation_matrix.sav'.
RENAME VARIABLES (@1=factor_1) (@2=factor_2) (@3=factor_3).
STRING ROWTYPE_ (a8) VARNAME_ (a8).
COMPUTE ROWTYPE_='CORR'.
COMPUTE VARNAME_='factor_1'.
IF $CASENUM=2 VARNAME_='factor_2'.
IF $CASENUM=3 VARNAME_='factor_3'.
EXECUTE.
```

If you have more than 3 factors, you have to add them here

*You then have to manually add a case to the data file. Insert this case before
*the other cases. It should have "N" for the ROWTYPE_ variable, nothing for the
*VARNAME_ variable, and for each of the factor_ variables, the N from the original
factor analysis. After you do this, run the SAVE command.

```
SAVE OUTFILE='C:\temp\Second_order_factor_input.sav'
/KEEP=ROWTYPE_ VARNAME_ factor_1 to factor_3.
```

*Before running the FACTOR command, you'll need to decide on the extraction
*and rotation methods and how many factors you want extracted if you don't
*want FACTOR to choose based on the eigenvalues greater than 1 default
*criterion. Replace the question marks with your choices. If you want FACTOR
*to choose the number of factors, remove the "FACTORS(?)" keyword from the
*CRITERIA subcommand. If you don't want the second order solution rotated,
specify NONE on the ROTATION subcommand.

```
FACTOR /MATRIX=IN(COR='C:\temp\Second_order_factor_input.sav')
/PRINT INITIAL EXTRACTION ROTATION
/CRITERIA FACTORS(1) ITERATE(50)
/EXTRACTION PAF
/ROTATION promax.
```

What is Cronbach's alpha (and what is it not)?

Cronbach's alpha (Cronbach, 1951) is one of the most widely reported indicators of scale reliability in the social sciences. It has some conveniences over other methods of indicating reliability, and it also has some drawbacks. It also has many misconceptions. There is even a test to determine if alpha is the same across two samples (Feldt, 1980), and others have proposed methods to compute confidence intervals for alpha (see Barnette, 2005).[49]

Let us start with the original goal for alpha. Prior to Cronbach's seminal work in this area, the reliability of a scale in a particular sample[50] was evaluated through methods such as test-retest correlations. This type of reliability is still discussed today in psychometrics textbooks, but has serious drawbacks. This can include the difficulty of convening the same group of individuals to re-take instruments, memory effects, and attenuation due to real change between administrations. Another serious drawback is particular to constructs (e.g., mood, content knowledge) that are expected to change over time. Thus, as Cronbach himself put it, test-retest reliability is generally best considered an index of *stability* rather than reliability *per se*.

The split-half reliability estimate was also developed early in the 20th century. To perform this evaluation, items are divided into two groups (most commonly, even and odd numbered items) and scored. Those two scores are then compared as a proxy for an immediate test-retest correlation. This too has drawbacks—the number of items is halved, there is some doubt as to whether the two groups of items are parallel, and different splits of items can yield different coefficients. The Spearman-Brown correction was developed to help correct for the reduction in item number and to give a coefficient intended to be similar to the test-retest coefficient. As Cronbach (1951) pointed out, this coefficient is best characterized as an indicator of equivalence between two forms, much as we today also talk about parallel forms.

Alpha and Kuder-Richardson coefficient of equivalence. The Kuder Richardson Formula 20 (KR-20) was developed to address some of the concerns over other forms of reliability, particularly split half, and preceded alpha. It is specific to items scored ether "0" or "1" as in many academic tests or scales such as the Geriatric Depression Scale we use as an example earlier in the book. Alpha is more general than KR20, but KR20 and alpha will arrive at the same solution if items are binary. Thus, it does not appear that KR-20 is necessary in modern statistical methodology.

The correct interpretation of alpha. Cronbach (1951) himself wrote and provided proofs for several assertions about alpha. These include:

- Alpha is n/n-1 times the ratio of inter-item covariance to total variance—in

---

9 AFTER THE EFA: INTERNAL CONSISTENCY

After you are done with the exploratory factor analysis, your journey is just beginning, rather than ending. The exploration phase might be drawing to a close, but the psychometric evaluation of an instrument is merely starting. The process of performing exploratory factor analysis is usually seeking to answer the question of whether a given set of items forms a coherent factor (or several factors). After we decide whether this is likely, evaluating how well those constructs are measured is important. Along the way, we can also ask whether the factor being examined *needs* all the items in order to be measured effectively.

To fully evaluate an instrument, we should evaluate whether the factors or scales that we derive from the EFA are reliable, confirmed in a new sample, and stable (invariant) across multiple groups. In this chapter, we will briefly look at the most common method of assessing scale reliability, Cronbach's alpha.

Let us first start with a discussion of the modern view of reliability and validity. When developing a scale to be used in research, there is a delicate dance between focusing on creating a scale that is a "good" scale, and the acknowledgment in modern research methods that things like factor structure, reliability, and validity are joint properties of a scale and of the particular sample data being used (Fan & Thompson, 2001; Wilkinson, 1999). It should be self-evident to modern researchers that a scale needs to be well-developed in order to be useful, and that we do that in the context of a particular sample (or series of samples, as we recommended when discussing replication). Thus, those of us interested in measurement must hold two somewhat bifurcated ideas in mind simultaneously- that a scale can be stronger or weaker, and that scales are only strong or weak in the context of the particular sample being used. This can lead to a nihilistic mindset if carried too far, so I recommend we take a moderate position in this discussion: that scales can be more or less strong, but that all scales need to be evaluated in the particular populations or data that they reside in.

---

[49] Although neither practice seems to have been adopted widely in the literature I am familiar with

[50] Back in the middle 20th century, reliability and validity was discussed as a property of the scale (i.e., the Osborne Obsequiousness Scale is reliable and valid). Modern APA and other guidelines recommend we talk about reliability and validity as the property of samples not instruments. However, some instruments do tend to appear more reliable across samples, and some less so. Hence the need for replication…

other words, a direct assessment of the ratio of error (unexplained) variance in the measure

- The average of all possible split half coefficients for a given test.[51]
- The coefficient of equivalence from two tests composed of items randomly sampled (without replacement) from a universe of items with the mean covariance as the test or scale in questions
- A lower-bound estimate of the coefficient of precision (accuracy of the test with these particular items) and coefficient of equivalency (simultaneous administration of two tests with matching items)
- The proportion (lower bound) of the test variance due to all common factors among the items

All of these lead me to conclude that the standards we use for alpha, and the average alphas found in strong journals, are not good enough. As Nunnally and Bernstein (1994, p. 235) point out distill from all this, alpha is an expected correlation between one test and an alternative form of the test containing the same number of items. The square root of alpha is also, as they point out, the correlation between the score on a scale and errorless "true scores." Let us unpack this for a moment.

This means that if one has an alpha of 0.80 for a scale that is interpreted as the expected correlation between that scale and another scale sampled from the same domain of items with the same covariance and number of items. The square root of 0.80 is 0.89, which represents an estimate of the correlation between that score and the "true scores" for that construct. As you probably know, the square of a correlation is an estimate of shared variance, so squaring this number leaves us back to the proportion of "true score" in the measurement (and 1- α is the proportion of error variance in the measurement).

A review of educational psychology literature from 1969 and 1999 indicated average (reported) alphas of 0.86 and 0.83, respectively (Osborne, 2008b). This is not bad, but keep in mind that even in modern, high-quality journals, only 26% of articles reported this basic data quality information. Despite these optimistic averages, it is not difficult to find journal articles in almost any discipline reporting analyses with alphas much lower. Poor measurement can have profound (and often unpredictable) effects on outcomes.

Taking the example of multiple regression, with each independent variable added to a regression equation, the effects of less than perfect reliability on the strength of the relationship becomes more complex and the results of the analysis more questionable. One independent variable with less than perfect reliability can lead to each subsequent variable claiming part of the error variance left over by the unreliable variable(s). The apportionment of the explained variance among the independent variables will thus be incorrect and reflect a mis-estimation of the true population effect. In essence, low reliability in one variable can lead to substantial over-estimation of the effect of another related variable. As more independent variables with low levels of reliability are added

---

[51] This implies that there is a distribution of split half coefficients based on different splits, and that alpha is the mean of all these splits. An interesting idea that many of us miss, as we focus just on the one number we calculate.

to the equation, the greater the likelihood that the variance accounted for is not apportioned correctly. Ultimately, some effects can end up masked (creating a Type II error), with other effects inflated inappropriately in the same analysis, potentially leading to Type I errors of inference (Osborne, 2013). Thus, one thesis of this chapter is that better measurement is preferable to less good measurement.[52]

What alpha is *not*. Although Note that α is *not* a measure of unidimensionality (an indicator that a scale is measuring a single construct rather than multiple related constructs) as is often thought (Cortina, 1993; Schmitt, 1996). Unidimensionality is an important assumption of α, in that scales that are multidimensional will cause α to be under-estimated if not assessed separately for each dimension, but high values for α are not necessarily indicators of unidimensionality (e.g., Cortina, 1993; Schmitt, 1996).

Factors that influence alpha

Average inter-item correlation. All other things being equal, alpha is higher when the average correlation between items is higher. But even Cronbach specifically pointed out that when inter-item correlations are low, alpha can be high with enough items with low intercorrelations. This is one of the chief drawbacks to interpretability of alpha- that with enough mostly unrelated items, alpha will move into the "reasonable" range that most researchers use as a rule of thumb.

Length of the scale. As mentioned above, all other things being equal, longer scales will have higher alphas.

Reverse coded items (negative item-total correlations). Many scales are constructed with reverse-coded items. However, alpha cannot provide accurate estimates when the analysis includes items with negative item-total correlations. Thus, any item that is expected to have a negative item-total correlation (e.g., if the factor loading is negative when most others are positive) should be reversed prior to analysis.

Random responding or response sets. Random responding (discussed earlier) tends to attenuate all of these estimates because it primarily adds random error. Thus, failure to identify this issue in your data will lead to under-estimation of the internal consistency of the data. Response sets can have a variety of effects, depending on the response set. Some types of response sets will inflate alpha estimates and some can attenuate alpha (for an overview of response sets, and how one can identify them, you might see Osborne & Blanchard, 2011)

Multidimensionality. The assumption of alpha is that all items within a particular analysis represent a single dimension, or factor. To the extent that assumption is

---

[52] Most people would agree this statement is "self-evident"- a nice way of saying "well, duh!" but it is surprising that this simple "well, duh!" sentiment is so problematic in practice…

violated, the estimate of alpha will be mis-estimated. Thus, the factor structure of the scale should be considered before submitting items to this type of analysis.

Outliers. Outliers (inappropriate values) usually have the effect of increasing error variance, which would have the effect of attenuating the estimate of alpha. Thus, data should be carefully screened prior to computing alpha.

Other assumptions of alpha. Alpha was built for a time when researchers often summed or averaged items on scales. Many researchers do this today. Of course, when summing or averaging items in a scale, you are making an assumption that all items contribute equally to the scale- that the weighting of each item is identical. Alpha also assumes that all items contribute equally. Yet from what we have seen in earlier chapters, that might not be a valid assumption. For example, in Chapter 2 we saw that the pattern loadings for the engineering problem solving items ranged from 0.79 to 0.91, and for the GDS the loadings ranged from 0.23 to 0.68 when only one factor was extracted. If you square the loadings to estimate the shared variance, this amounts to a range of 0.62 to 0.82 (for engineering problem solving) and from 0.05 to 0.46 for GDS.

Historically, this led to researchers creating "factor scores" which weighted each item by the factor loading to more closely approximate that a latent variable score might be. There is a whole chapter coming on why this is *not* a good idea. More modern measurement methods (IRT, Rasch) account for this more directly, and latent variable modeling (structural equation modeling, for example) also directly addresses this issue when estimating individual scores on latent variables.

There is not currently a good way to deal with this issue (in my opinion). Rasch and IRT have different methods of estimating reliability of measures, and in CFA/SEM there are also ways to assess goodness of fit, but those interested in estimating internal consistency via alpha must live with the violation of this assumption.

What is "good enough" for alpha?

Many authors have asserted that an alpha of 0.70 or 0.80 represent "adequate" and "good" reliability, respectively (e.g., Nunnally & Bernstein, 1994). Let us just say for the present that it is a continuous variable, and thus, higher is better, probably with diminishing returns once one exceeds 0.90 (which still represents about 10% error variance in the measurement).

What constitutes "good enough" also depends on the purpose of the data, and the method of analysis. Better data is always better, of course, but using the data to choose children for an educational program is different than evaluating correlations between constructs for a dissertation. And use of modern measurement (e.g., Rasch or IRT measurement) and modern analysis techniques (e.g., structural equation modeling) can help improve the situation.

Would error-free measurement make a real difference?

To give a concrete example of how important good measurement is, we can use the example from my survey of the Educational Psychology literature from 1998 to 1999

(Osborne, 2008b). This survey consisted of recording all effects from all quantitative studies published in the Journal of Educational Psychology (usually considered one of our top empirical journals) during the years 1998-1999.

Studies from these years indicate a mean effect size (d) of 0.68, with a standard deviation of 0.37. When these effect sizes are converted into simple correlation coefficients via direct algebraic manipulation (formulae and conversions derived from information in Cohen, 1988), $d = .68$ is equivalent to $r = .32$. Effect sizes one standard deviation below and above the mean equate to rs of .16 and .46, respectively (we will use these to represent "small" and "large" effects in the example below).

From the same review of the literature, where reliabilities (Cronbach's α) are reported, the average reliability is about $\alpha = .80$, with a standard deviation of .10.

Table 9.1 demonstrates what effects researchers might be expected to find if they had error-free measurement under a variety of scenarios. In all three columns, we can see that observed effects are substantially under-estimated. For example, looking at an average observed effect ($r = 0.32$), we can see that even if reliability was good ($\alpha = .90$), the effect is under-estimated by 30%. With "good" reliability of about $\alpha = 0.80$ we can see this effect is under-estimated by 60%, and by over 100% if alphas were 0.70 (which is often considered acceptable in top journals).

Table 9.1
The effects of imperfect measurement from Educational Psychology literature.

	And the observed effect was:		
If measurement was:	Small effect ($r = .16$, $r^2 = .025$)	Average effect ($r = .32$, $r^2 = .10$)	Large effect ($r = .46$, $r^2 = .21$)
Poor ($\alpha = .70$)	$r = .23$ $r^2 = .052$	$r = .46$ $r^2 = .21$	$r = .66$ $r^2 = .43$
Average ($\alpha = .80$)	$r = .20$ $r^2 = .040$	$r = .40$ $r^2 = .16$	$r = .58$ $r^2 = .33$
Good ($\alpha = .90$)	$r = .18$ $r^2 = .032$	$r = .36$ $r^2 = .13$	$r = .51$ $r^2 = .26$

Note: these data are adapted from those originally published in Osborne (2008b)

An example from my research. I now review an example from my personal research on identification with academics (self-concept *vis a vis* school). In one study a long while ago, I administered two closely-related scales (the School Perceptions Questionnaire and the Identification with School questionnaire, (Osborne, 1997; Voelkl, 1997) to high school students. Alphas were calculated to be $\alpha = 0.88$ and 0.81, respectively. These levels are widely considered "good." After averaging the items to create composite scores and testing assumptions, the simple correlation[53] was calculated

---

[53] You didn't expect the guy writing the book on data cleaning to skip that part, did you? All assumptions met.

to be $r = 0.75$, which translates to a coefficient of determination (% variance accounted for) of 0.56. This is generally a pretty strong correlation for the social sciences, and is reasonable considering these are two measures of similar constructs. The corrected correlation (formulae available in many places including (Cohen, Cohen, West, & Aiken, 2002) is calculated to be $r = 0.89$, which would represent a coefficient of determination of 0.79. If we assume that this corrected effect size of 0.79 is the correct population estimate, the original correlation between two measures with "good" reliability lost almost 30% of the effect size.

Although we cannot directly know the population correlation for this example, we can simulate what perfect measurement might yield as a correlation between these two variables using AMOS structural equation modeling software to construct latent variables representing each of these scales. While structural equation modeling is a relatively advanced procedure, and getting into the intricacies of the analysis is beyond the scope of this chapter, for our purposes all you need to understand is that SEM can be used to estimate relationships between variables as though they were measured perfectly. The estimate, therefore, of the correlation under perfect correlation was $r = 0.90$ (coefficient of determination of 0.81), very close to the calculated corrected correlation effect size.

Unfortunately, while alphas in the 0.80 range are common, published research based on data with much lower reliabilities are also not difficult to find. In fact, it is not difficult to find alphas under 0.70, despite the fact that means that a substantial portion of the effect size is lost! This should be considered an undesirable state of affairs in the 21st century, particularly when it is relatively simple to improve measurement in scales by increasing item numbers or through analysis by using modern methods like structural equation modeling.

Sample size and the precision/stability of alpha-empirical confidence intervals

Fan and Thompson (2001) point out that few authors provide context in their papers as to the precision of their effect size point estimates. But alpha is merely a point estimate like any other statistic that we have talked about thus far. Alpha is also sample-dependent, with representative samples better than biased samples, and larger samples better than smaller samples. When authors report alpha without the context of confidence intervals, readers have no way to understand how precise that estimate is, and how likely that alpha is to replicate. As we briefly discussed earlier in the chapter, there have been attempts to construct methods to calculate CIs for alpha, but these have not gained traction in routine practice. However, with bootstrapping, we can easily provide empirical estimates that are valuable for readers.

Marsh SDQ. First, let's start with some exploration in the variability of alphas as a function of effect size. Let us return to the Marsh SDQ example we have been using throughout the book. The parent subscale is composed of 5 items;

(Par1) My parents treat me fairly
(Par2) I do not like my parents very much

(Par3) I get along well with my parents
(Par4) My parents are usually unhappy or disappointed with what I do
(Par5) My parents understand me

Obviously PAR2 and PAR4 are reversed in direction. If you calculate alpha using these five items as is you will see a negative alpha (which is impossible) and some sage advice from SPSS (as you can see in Table 9.2):

Table 9.2
Initial alpha is negative from reverse coded items

Reliability Statistics

Cronbach's Alpha[a]	N of Items
-.338	5

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Par1	13.41	4.864	.194	-.834[a]
Par2	16.48	8.945	-.417	.195
Par3	13.45	4.791	.175	-.829[a]
Par4	16.03	8.790	-.406	.220
Par5	13.89	4.188	.115	-.873[a]

a. The value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings.

Table 9.3
Final alpha statistics

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Par1	19.1763	19.792	.643	.801
Par2r	18.9161	19.525	.596	.812
Par3	19.2172	18.575	.733	.776
Par4r	19.3684	19.075	.569	.820
Par5	19.6597	17.206	.659	.796

Once the items are recoded so that they all have positive intercorrelations, $\alpha = 0.834$ with an $N = 15661$. Further, the item-total correlations all range from 0.569 to 0.733, which is reasonable. If we saw very low statistics in that column, we could examine whether the scale was appropriate for modification/removal of that item.

Let's start with this as our gold standard "population" statistic, and see what small samples can do to the estimation of alpha. I have seen alpha estimated on samples much smaller than 50, but will use that as our "small sample" example. I had SPSS randomly pull 1000 samples of $N = 50$ each, and compute alpha for the same scale.
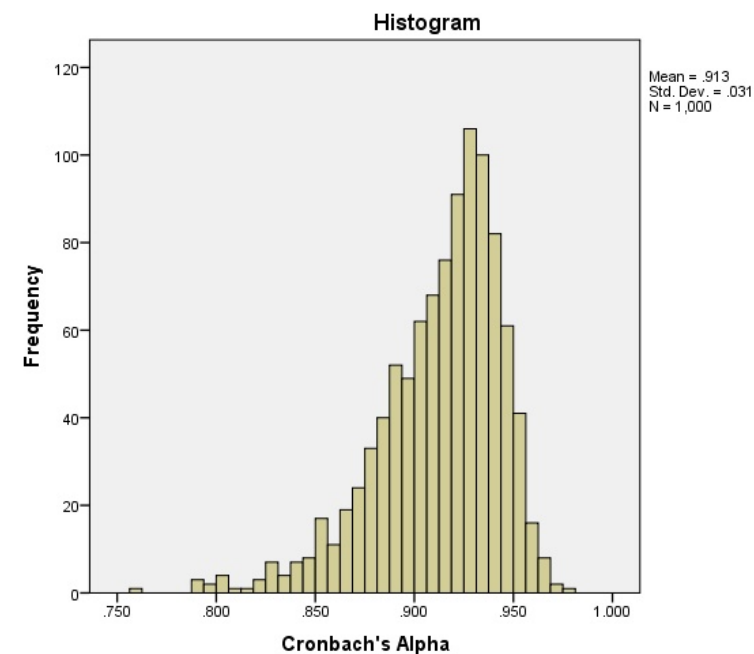
Figure 9.1: Distribution of Cronbach's alpha with samples of N=50
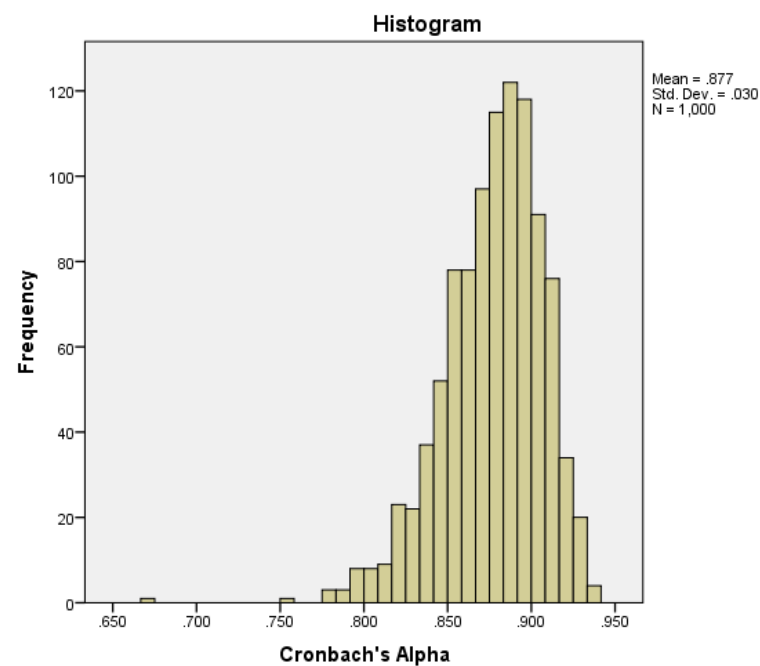

Figure 9.2: Distribution of Cronbach's alpha with samples of N=100

As you can see in Figure 9.1, with such small samples, alpha became somewhat volatile. Interestingly, the mean alpha of the 1000 samples was 0.91, much higher than the "population" mean of $\alpha = 0.83$. The range of observed alphas was from a low of $\alpha = 0.76$ to a high of $\alpha = 0.98$. Thus, with small samples, even a scale such as this (which has reasonably strong psychometric properties) had a good deal of volatility. In particular, the small sample alphas seemed to tend toward significant over-estimation of alpha.
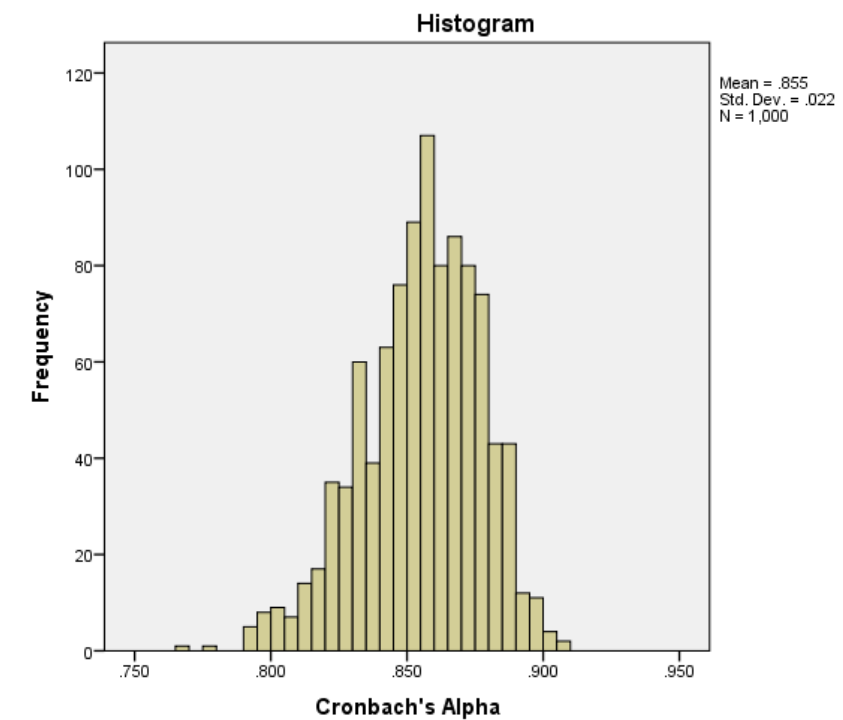

Figure 9.3: Distribution of Cronbach's alpha with samples of N=250

Doubling the sample size to $N = 100$ helped narrow range somewhat, but still allowed for some troubling outliers, as you can see in Figure 9.2. The minimum was 0.673, and the maximum was 0.938. However, the mean was 0.88, which was closer to the "population" alpha. Even at $N = 250$, the distribution of samples has a broad range (from $\alpha = 0.77$ to 0.91, with a mean of $\alpha = 0.855$, but a much smaller standard deviation, as you can see in Figure 9.3).

Geriatric Depression Scale. You may remember that the GDS gave us some problems when attempting to perform EFA, as it was unclear whether it was a single factor or multiple factors. Let us assume that we decided it was a single factor (remember that alpha is not a test of unidimensionality!). If we calculate alpha for the GDS, we estimate $\alpha = 0.889$ and in interesting list of item statistics (presented in Table 9.4). If you examine the item-total correlations, you will see a wide range of correlations, ranging between 0.256 and 0.615. What this seems to indicate is that, if

this is used as a single scale, there are items that could be deleted with little loss to reliability of measurement.

Sample size also influenced the instability of alpha estimates for this scale. With 1000 random samples of $N = 50$ each examined, the range of alpha was 0.75 to 0.95, with a mean of 0.89. Doubling the sample size to $N = 100$ in this case did not produce a significant improvement. The range of alpha across 1000 samples was $\alpha = 0.74$ to $\alpha = 0.93$ (mean $\alpha = 0.87$). Increasing the sample size to $N = 250$ in this case decreased the variability a bit, with 100 samples of this size ranging from $\alpha = 0.80$ to $\alpha = 0.91$ (mean $\alpha = 0.87$).

Table 9.4
Item-total statistics for the GDS

Item-Total Statistics				
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
GDS01	5.34	27.593	.523	.885
GDS02	5.21	27.414	.403	.887
GDS03	5.37	27.739	.545	.885
GDS04	5.27	27.217	.504	.885
GDS05	5.33	27.723	.456	.886
GDS06	5.28	27.229	.512	.884
GDS07	5.39	28.364	.401	.887
GDS08	5.37	28.297	.363	.887
GDS09	5.35	27.608	.540	.884
GDS10	5.32	27.241	.575	.883
GDS11	5.21	27.386	.411	.887
GDS12	5.14	27.411	.365	.888
GDS13	5.25	27.645	.380	.887
GDS14	5.30	28.299	.256	.889
GDS15	5.36	28.090	.415	.887
GDS16	5.32	27.116	.612	.883
GDS17	5.32	27.169	.615	.883
GDS18	5.38	28.290	.388	.887
GDS19	5.12	26.505	.549	.883
GDS20	5.16	27.081	.446	.886
GDS21	4.99	26.747	.464	.886
GDS22	5.39	28.180	.479	.886
GDS23	5.36	28.211	.373	.887
GDS24	5.23	27.547	.382	.887
GDS25	5.35	27.715	.521	.885
GDS26	5.20	27.067	.478	.885
GDS27	5.25	27.726	.355	.888
GDS28	5.22	27.369	.421	.886
GDS29	5.20	27.699	.329	.889
GDS30	4.97	27.244	.363	.889

Does bootstrapping small samples provide valuable information?

Ideally, bootstrap analyses of small samples would include the population parameter in the 95% confidence interval, and provide information about the precision of the estimate (and thus the potential replicability of the statistic). To explore this, we return to the SDQ parent subscale (with reverse coded items). I randomly selected a sample of $N = 100$ from the "population" of 15,661. The alpha in this small sample was a good estimate of the population: $\alpha = 0.84$. Analysis of 5000 bootstrapped samples revealed an empirical 95%CI of [0.77, 0.89], which does include the population parameter. The broad span of the CI suggests other similar samples would see variation but all within a reasonable range.

Larger samples produce more precise estimates and narrower confidence intervals. A bootstrap resampling analysis of a different random sample of $N = 500$ (see Figure 9.4) produced a 95%CI of [0.80, 0.86].
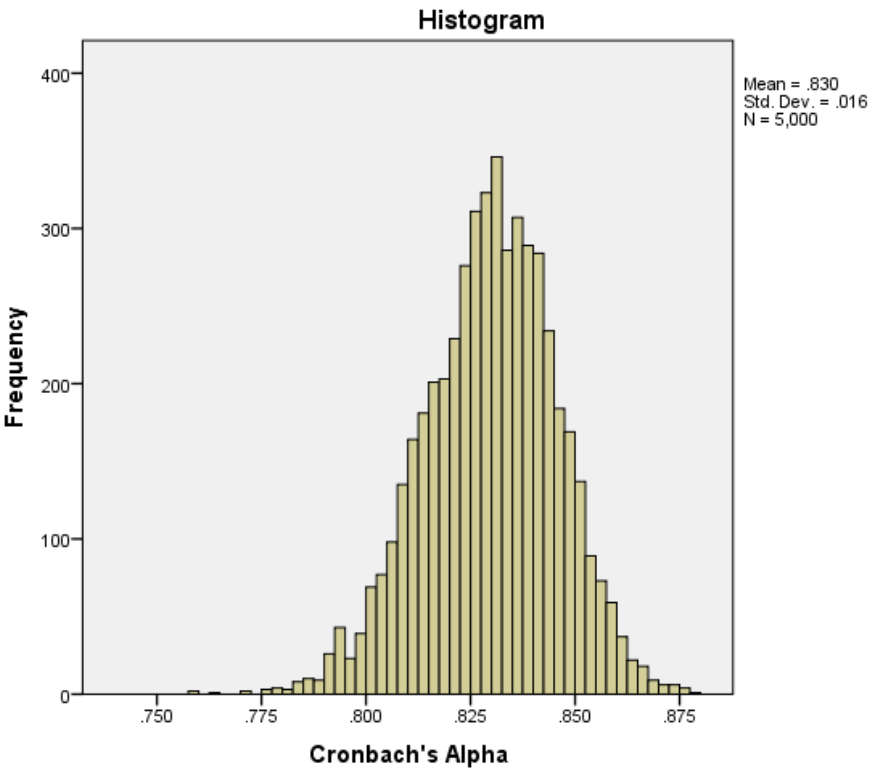

Mean = .830
Std. Dev. = .016
N = 5,000

Figure 9.4: Bootstrap resampling of the Marsh SDQ data set, N=500 each.

Bootstrap analysis of item-total correlations. We can bootstrap many different statistics, such as item-total correlations. Table 9.5 presents bootstrapped estimates of the first 10 items from the GDS data using 5000 samples. There is substantial variation (as presented in Figure 9.5) and some broad confidence intervals (e.g., item 8), and some narrower CIs (e.g., item 10). I have purposely left the rest of the table blank so

you can have the enjoyment of playing with this analysis. Your challenge is to fill in the rest (or at least some) of the table.

Table 9.5
Bootstrap analysis of GDS the first 10 item-total correlations.

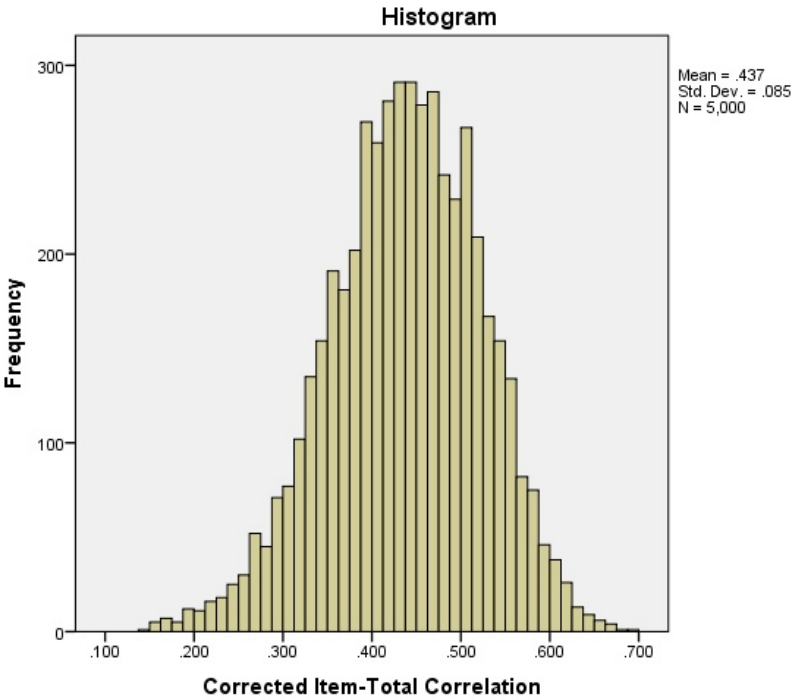|	Average item-total correlation	95% CI
GDS01	0.44	0.26, 0.59
GDS02	0.37	0.23, 0.50
GDS03	0.50	0.33, 0.65
GDS04	0.50	0.36, 0.63
GDS05	0.48	0.32, 0.62
GDS06	0.44	0.29, 0.58
GDS07	0.19	0.04, 0.35
GDS08	0.27	0.06, 0.46
GDS09	0.43	0.25, 0.58
GDS10	0.64	0.52, 0.74

Figure 9.5: Item-total correlation from GDS item #1 bootstrap analysis

Chapter 9 Summary

Cronbach's alpha is one of the most widely-reported statistics relating to reliability in the social sciences. However, it is still not reported in a majority of articles in modern, high-quality research journals. This might stem from the anachronistic assertion that scales are more or less reliable. The modern view of reliability is that it is a property of the data and sample, not the scale or test, and as such, should be evaluated and reported in every study where it applies.

Alpha is interpreted as an estimate of all possible split-half statistics, and can also be interpreted as the percent of variance that is "true score" variance. Thus, if you have a measure with $\alpha = 0.80$, there is about 80% that is "true score" and about 20% error in the measurement. Of course, this is both a classical test theory view, and an estimate, since we just demonstrated that alpha can vary widely across samples (particularly small samples).

This variability across samples is not something often discussed in the psychometrics literature, but it is *important*. This highlights the view that reliability is a property of the sample, and should be attended to and reported in each study. Additionally, we reviewed an application of bootstrap analysis for estimation of 95% confidence intervals around alpha.

Alternatives to alpha. Alpha is the most widely-accepted evaluation of reliability (internal consistency) in the modern literature, despite the fact that most papers in top journals fail to report reliability information. Alpha is over sixty years old, however, and in the meanwhile other alternatives have been developed. Rasch modeling, for example, produces interesting information about reliability from that perspective, as does Item Response Theory analyses. Structural equation modeling (SEM) allows us to explicitly model the latent construct and directly analyze it, eliminating the need for alpha altogether. Where possible, it seems desirable to utilize modern measurement methodologies.

Chapter 9 exercises

1. Describe the conceptual meaning of Cronbach's alpha. If your scale has $\alpha = 0.85$, how do you interpret that number?
2. Download the $N = 100$ and $N = 500$ random samples of the SDQ data from the book web site and replicate the bootstrapped 95%CIs for alpha.
3. Using another data set from the book or your own data:
    a. Make sure all items are coded in the same direction (recode items where necessary) and calculate alpha. Also examine the item-total correlations to see if any items could be removed from the scale to improve reliability.
    b. Bootstrap various size samples to see how confidence intervals around alpha become wider or narrower with sample size.
    c. Interpret alpha in terms of quality of measurement, and also discuss the results of bootstrap analysis.
4. Explore bootstrap analysis of item-total correlations on the GDS data, replicating the first 10 items and filling in the rest. Remember, your data will vary in small ways as you resample because of the random nature of the resampling. However, the overall distributions and means should be close within a few decimal points.

10 SUMMARY AND CONCLUSIONS

This journey started about a decade ago when one of my students walked into my office and shared with me her frustration over conflicting advice and directives from different doctoral committee members. We talked for a while about exploratory factor analysis during that meeting, and I ended up deciding that there was not a clear literature on the issue, but that the issue at hand was an empirical question that we could answer together by running a bunch of simulations. Our initial discussions and commitment to find empirically driven guidelines led to two articles and several conference presentations on best practices related to EFA, one of which (Costello and Osborne, 2005) has been cited about 2700 times as I write this.

It has been a fun and unexpected journey. I never set out to focus on EFA, but it is such a confounding, controversial, and mis-used technique that it has provided lots of fun and fodder for this type of endeavor. After publishing each article, I routinely assumed that was the last time I would write on factor analysis. It was never my goal to focus so much effort on EFA throughout a good portion of my career. Then, after publishing my most recent book on logistic regression, I was mulling over the direction I would take for my next project. I was leaning in a completely different direction – and still am—but woke up one morning and realized I needed to write this book first. I don't know why, but it has been fun bringing this decade-long strand of work to fruition in a single place, and adding new perspectives (bootstrap analysis, for example).

I hope you agree it has been worthwhile. My goal is to collect and elaborate on the lessons learned over the recent decade, and to put them in a single place that is easily accessible. For those of you who have persevered and have reached this part of the book, I hope that you have drawn the following conclusions:

1. **Keep the "E" in EFA!** Many researchers have attempted to perform confirmatory analyses by performing exploratory analyses. Many researchers use confirmatory language and draw confirmatory conclusions after performing exploratory analyses. This is not appropriate. EFA is a fun and important technique, but it is what it is. We need to remember to honor that and use confirmatory techniques when we desire to draw those types of conclusions.

2. **EFA is a large sample technique.** I hope that through the course of the book you have become convinced that the best results from EFA come when the sample is appropriately large. A starting point for that is a sample that includes 20 cases for each variable in the analysis—and that is what I would consider the minimum, not the maximum, if you want a robust, generalizable result. I have had students and colleagues show me analyses that had fewer cases than variables. That is never a good state of affairs, in my opinion.

3. **Useful results are those that are precise and generalizable.** In my mind, the most useful results are those that we can generalize to other samples, or use to draw good inferences about the population as a whole. In my mind, the worst use of anyone's time is to publish or present results that are not replicable, or are so imprecise that we cannot draw any conclusions about anything other than the individuals in the original sample. Large samples and clean data (in addition to strong factor loadings and larger numbers of strongly-loading variables per factor) contribute to this mission. Small samples and weak loadings (and few variables per factor) make for messy, conflicting, and useless results.

4. **Principal Components Analysis is not Exploratory Factor Analysis.** I cannot tell you how tired I am of this debate, and of those who insist there is a use for PCA. Honestly, I don't really care which side of the debate you are on. If you feel some compelling reason to use PCA, then I hope this book can guide you as well. Most of the best practices we have covered in this book also apply to PCA. If you insist on using PCA, at least do it with large samples, clean data, and with the limitations of the procedure clearly and overtly admitted. ANOVA has limitations too, but I use it on occasion, when appropriate, and with best practices in mind.

5. **If you use EFA, don't use the defaults!** If you want to consider yourself to be modeling and exploring latent variables in the best way possible, you want to use ML or PAF extraction (depending on whether your data meets the assumptions of ML), and I think you want to use oblique rotation (either Oblimin or Promax seems to work fine in most cases—if one doesn't work, try the other). Scholars in this area spend so much energy arguing about which extraction or rotation technique is best. But keep mantra in mind- this is just an exploration. Thus, it is a low-stakes endeavor. Whatever you find from EFA has to be confirmed in a large sample confirmatory analysis. With this in mind, all this arguing seems to be a bit of a waste of effort, in my opinion.

6. **Use multiple decision rules when deciding how many factors to extract.** Another point of constant argument in this field seems to be what decision rule is best in guiding someone on how many factors to extract. We reviewed several, and none are perfect. Just in our three examples, one had a clearly uninterpretable scree plot, one parallel analysis produced what I consider to be questionable guidance, and one MAP analysis that was clearly (to my eye, anyway) ambiguous. The other criteria were also at times confusing and problematic. The best guide is theory, and beyond that, whatever provides the results that *make the most sense*. If you cannot make sense of the results—in other words, if you cannot easily explain to someone what each factor means—then you need to go back to exploring. Because any model you produce has to be

confirmed with CFA in the context of a new sample, this seems to me the most sensible approach. Thanks to Brian O'Connor, we have easily accessible ways of trying out modern decision criteria (MAP, Parallel analysis). Use them, but no one decision rule will be perfect in all situations.

7. Replicate your results. If you have two good samples, you can present replication statistics like I reviewed in Chapter 4, or you can put a single sample to work in bootstrap analysis, like I explored in Chapter 5. It's not easy nor is it automatic, but with the syntax/macro I share, it is not impossible. And I think that it provides invaluable perspective on your results. I wish this mandate to replicate results would permeate every research lab, regardless of what statistical techniques they use. The lessons contained in these chapters are equally valid if you are performing ANOVA or regression analyses, hierarchical linear modeling, or nonparametric techniques. Replicate your results, bootstrap your analyses, and report (and interpret) confidence intervals for important effects so we, as readers, can get more out of the hard work you put into your research.

8. Clean your data, and deal with missing data appropriately. Garbage in, garbage out. I won't belabor this point- but I hope you take it seriously. If I don't see you address whether you checked your data, tested assumptions, and dealt appropriately with missing data, I am going to be wondering whether anything else you report matters.

9. **Have fun!** The ability and training to perform research like this is a wonderful gift. I have been lucky enough to spend the last twenty-five years doing quantitative research, primarily in the social sciences, and I have enjoyed every minute of it. Those of us who perform data analysis[54] are the ones who are present at the moment each tiny bit of knowledge is created. We create knowledge- we ask questions and find answers. Sometimes those answers are not what we expect, which is an opportunity to ask better questions or learn something unexpected. I cannot think of a more rewarding way to spend my career, and I hope each one of you experiences the same joy and thrill from your research.

Thank you for taking time to read my work. I always welcome feedback or communication from readers. The best way to reach me is through email at: jasonwosborne@gmail.com. I hope you find the ancillary materials I will put, and will continue to develop, on my website (http://jwosborne.com) useful. Happy researching!

---

[54] My respected colleagues who perform qualitative research are included in this generalization here. What I find so compelling is the process of analyzing data, not the mode of analysis.

ABOUT THE AUTHOR

Jason is currently Professor and Chair of Educational and Counseling Psychology, Counseling, and College Student Personnel at the University of Louisville in Louisville, Kentucky (USA). He has been named an Accredited Professional Statistician™ by the American Statistical Association. This is his fifth book on best practices in quantitative methods, and he has published over 70 articles and presented scores of times at national and international conferences. Information about his books, as well as ancillary materials (data sets, syntax examples, how-to guides, answer keys, errata, etc.) can be found on his web site: http://jwosborne.com.

The journey continues!

REFERENCES

Aleamoni, L. M. (1976). The relation of sample size to the number of variables in using factor analysis techniques. *Educational and psychological measurement, 36*, 879-883.

Baggaley, A. R. (1983). Deciding on the ratio of number of subjects to number of variables in factor analysis. *Multivariate Experimental Clinical Research, 6*(2), 81-85.

Barnette, J. J. (2005). ScoreRel CI: An Excel program for computing confidence intervals for commonly used score reliability coefficients. *Educational and psychological measurement, 65*(6), 980-983.

Barrett, P. (1986). Factor comparison: An examination of three methods. *Personality and Individual Differences, 7*(3), 327-340.

Barrett, P. T., & Kline, P. (1981). The observation to variable ratio in factor analysis. *Personality study and group behavior, 1*, 23-33.

Beach, D. A. (1989). Identifying the Random Responder. *Journal of Psychology, 123*(1), 101.

Bentler, P., & Kano, Y. (1990). On the equivalence of factors and components. *Multivariate Behavioral Research, 25*(1), 67-74.

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using self-report methodology. *Psychological assessment, 4*, 340-345.

Bobko, P., & Schemmer, F. M. (1984). Eigen value shrinkage in principal component based factor analysis. *Applied Psychological Measurement, 8*, 439-451.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Bond, T. G., & Fox, C. M. (2006). *Applying the Rasch model: Fundamental measurement in the human sciences*: Psychology Press.

Bovaird, J. A. (2003). *New applications in testing: Using response time to increase the construct validity of a latent trait estimate*. (64), ProQuest Information & Learning, US. Retrieved from http://www.lib.ncsu.edu/cgi-bin/proxy.pl?server=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2003-95016-048&site=ehost-live&scope=site

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality, 54*(1), 106-148. doi: 10.1111/j.1467-6494.1986.tb00391.x

Byrne, B. M. (2010). *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming*. New York, NY: Routledge.

Carroll, J. B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika, 18*(1), 23-38.

Cattell, R. B. (1965). A Biometrics Invited Paper. Factor Analysis: An Introduction to Essentials I. The Purpose and Underlying Models. *Biometrics, 21*(1), 190-215. doi: 10.2307/2528364

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245-276.

Cliff, N. (1970). The relation between sample and population characteristic vectors. *Psychometrika, 35*, 163-178.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (second edition)*. Hillsdale, NJ: Lawrrence Erlbaum.

Cohen, J., Cohen, P., West, S., & Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ: Lawrence Erlbaum.

Cole, J. C. (2008). How to Deal With Missing Data. In J. W. Osborne (Ed.), *Best Practices in Quantitative Methods*. Thousand Oaks, CA: Sage Publishing.

Comfrey, A. L., & Lee, H. B. (1992). *A First Course in Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cortina, J. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-98.

Costello, A. B., & Osborne, J. W. (2005). Exploratory Factor Analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(7), 1-9.

Cronbach, L. J. (1942). studies of acquiescence as a factor in the true-false test. *Journal of educational Psychology, 33*, 401-415.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and psychological measurement, 10*, 3-31.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Crowne, D., & Marlowe, D. (1964). *The approval motive*. New York: Wiley.

Curtin, T., Ingels, S., Wu, S., & Heuer, R. (2002). NELS 1988/2000: Base year to fourth follow-up data user's manual. *Washington, DC: National Center for Education Statistics*.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 189-212.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation, 14*(20), 1-11.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*, 5-18.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap* (Vol. 57): Chapman & Hall/CRC.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272.

Fan, X., & Thompson, B. (2001). Confidence Intervals for Effect Sizes Confidence Intervals about Score Reliability Coefficients, Please: An EPM Guidelines Editorial. *Educational and psychological measurement, 61*(4), 517-531.

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika, 45*(1), 99-105.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment, 7*(3), 286.

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology, 39*, 291-314.

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology, 39*(2), 291-314.

Forina, M., Armanino, C., Lanteri, S., & Leardi, R. (1989). Methods of varimax rotation in factor analysis with applications in clinical and food chemistry. *Journal of Chemometrics, 3*(S1), 115-125. doi: 10.1002/cem.1180030504

Goodfellow, L. D. (1940). The human element in probability. *The Journal of General Psychology, 33*, 201-205.

Gorsuch, R. L. (1990). Common factor analysis versus component analysis: Some well and little known facts. *Multivariate Behavioral Research, 25*(1), 33-39.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of personality assessment, 68*(3), 532-560.

Gorsuch, R. L. (1983). *Factor Analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430.

Guadagnoli, E., & Velicer, W. F. (1988). relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265-275.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Thousand Oaks, CA US: Sage Publications, Inc.

Hatcher, L. (1994). *A Step-by-Step Approach to Using the SAS® System for Factor Analysis and Structural Equation Modeling.* Cary, N.C.: SAS Institutte, Inc.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research. *Educational and psychological measurement, 66*(3), 393-416.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185.

Horst, P. (1941). A non-graphical method for transforming an arbitrary factor matrix into a simple structure factor matrix. *Psychometrika, 6*(2), 79-99. doi: 10.1007/BF02292176

Humphreys, L. G., Ilgen, D., McGrath, D., & Montanelli, R. (1969). Capitalization on chance in rotation of factors. *Educational and psychological measurement, 29*(2), 259-271.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 user's reference guide*: Scientific Software.

Kaiser, H., Hunka, S., & Bianchini, J. (1971). Relating factors between studies based upon different individuals. *Multivariate Behavioral Research, 6*(4), 409-422.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*(3), 187-200. doi: 10.1007/BF02289233

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement.*

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*(4), 401-415.

Kane, S. T. (2008). Minimizing Malingering and Poor Effort in the LD/ADHD Evaluation Process. *ADHD Report, 16*(5), 5-9.

Killeen, P. R. (2008). Replication Statistics. In J. W. Osborne (Ed.), *Best Practices in Quantitative Methods* (pp. 103-124). Thousand Oaks CA: Sage.

Kuncel, N. R., & Borneman, M. J. (2007). Toward a New Method of Detecting Deliberately Faked Personality Tests: The use of idiosyncratic item responses. *International Journal of Selection & Assessment, 15*(2), 220-231. doi: 10.1111/j.1468-2389.2007.00383.x

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the Number of Factors to Retain in EFA: an easy-to-use computer program for carrying out Parallel Analysis. *Practical Assessment, Research & Evaluation, 12*(2), 1-11.

Loehlin, J. C. (1990). Component analysis versus common factor analysis: A case of disputed authorship. *Multivariate Behavioral Research, 25*(1), 29-31.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, 38*(1), 88-91.

Lorge, I. (1937). Gen-like: Halo or reality? *Psychological Bulletin, 34*, 545-546.

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin, 109*(3), 502.

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. (1990). A multidimensional, hierarchical model of self-concept: Theoretical and empirical justification. *Educational Psychology Review, 2*(2), 77-172.

Marsh, H. W. (1994). Using the National Longitudinal Study of 1988 to evaluate theoretical models of self-concept: The Self-Description Questionnaire. *Journal of educational Psychology, 86*(3), 439.

McArdle, J. (1990). Principles versus principals of structural factor analyses. *Multivariate Behavioral Research, 25*(1), 81-87.

McCrae, R. R., & Costa Jr, P. T. (1997). Personality trait structure as a human universal. *American psychologist, 52*(5), 509.

Meier, S. T. (1994). *The Chronic Crisis in Psychological Measurement and Assessment: A Historical Survey.* San Diego, CA: Academic Press.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8*, 72-87.

Messick, S. (1991). Psychology and methodology of response styles. In R. E. Snow & D. E. Wiley (Eds.), *improving inquiry in social science* (pp. 161-200). Hillsdale, N.J.: Erlbaum.

Mulaik, S. A. (1990). Blurring the distinctions between component analysis and common factor analysis. *Multivariate Behavioral Research, 25*(1), 53-59.

Murphy, K. R., & Davidshofer, C. O. (1988). *Psychological testing.* Englewood Cliffs, NJ: Prentice Hall.

Nunnally, J. (1978). Psychometric methods. *New York: McGraw.*

Nunnally, J. C., & Bernstein, I. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw Hill.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior research methods, instruments, & computers, 32*(3), 396-402.

Osborne, J. W. (1997). Identification with academics and academic success among community college students. *Community College Review, 25*(1), 59-67.

Osborne, J. W. (2000). Prediction in Multiple Regression. *Practical Assessment, Research & Evaluation, 7*, n2.

Osborne, J. W. (2008a). Creating valid prediction equations in multiple regression: Shrinkage, Double Cross-Validation, and Confidence Intervals around prediction. In J. W. Osborne (Ed.), *Best practices in quantitative methods.* (pp. 299-305). thousand oaks, CA: Sage Publishing.

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*(2), 1 - 10.

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 Publications.

Osborne, J. W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology, 2*, 12. doi: 10.3389/fpsyg.2010.00220

Osborne, J. W., & Costello, A. B. (2004). Sample size and subject to item ratio in principal components analysis. *Practical Assessment, Research & Evaluation, 9*(11), 8.

Osborne, J. W., Costello, A. B., & Kellow, J. T. (2008). Best Practices in Exploratory Factor Analysis. In J. W. Osborne (Ed.), *Best Practices in Quantitative Methods* (pp. 205-213). Thousand Oaks, CA: Sage Publishing.

Osborne, J. W., & Fitzpatrick, D. C. (2012). Replication Analysis in Exploratory Factor Analysis: What it is and why it makes your analysis better. *Practical Assessment, Research & Evaluation, 17*(15), 2.

Osborne, J. W., & Jones, B. D. (2011). Identification with Academics and Motivation to Achieve in School: How the Structure of the Self Influences Academic Outcomes. *Educational Psychology Review, 23*, 131-158.

Pedhazur, E. J. (1997). *Multiple Regression in Behavioral Research: Explanation and Prediction.* Fort Worth, TX: Harcourt Brace College Publishers.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and data analysis methods.* (Vol. 1). thousand oaks, CA: Sage Publications.

Ray, C. L. (2009). The Importance of Using Malingering Screeners in Forensic Practice. *Journal of Forensic Psychology Practice, 9*(2), 138-146.

Rogers, R. (1997). Introduction. In R. Rogers (Ed.), *Clinical assessment of malingering and deception.* New York:: Guilford.

Schafer, J. (1997). *Analysis of incomplete multivariate data*: Chapman & Hall/CRC.

Schafer, J. (1999). Multiple imputation: a primer. *Statistical methods in medical research, 8*(1), 3.

Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological assessment, 8*(4), 350-353.

Smith, E. V., & Smith, R. M. (2004). *Introduction to Rasch Measurement.* Maple Grove, MN: JAM press.

Spearman, C. (1904). " General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology, 15*(2), 201-292.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences*: Lawrence Erlbaum.

Strong, E. K. J. (1927). A vocational interest test. *Educational record, 8*, 107-121.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics* (4th ed.). New York:: Harper Collins.

ten Berge, J. M. F. (1986). Rotation to perfect congruence and the cross validation of component weights across populations. *Multivariate Behavioral Research, 21*(1), 41-64.

ten Berge, J. M. F. (1996). The Kaiser, Hunka and Bianchini factor similarity coefficients: a cautionary note. *Multivariate Behavioral Research, 31*(1), 1-6.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *The Journal of Experimental Educational*, 361-377.

Thompson, B. (1999). Five Methodology Errors in Educational Research: The Pantheon of Statistical Significance and Other Faux Pas. In B. thompson (Ed.), *Advances in Social Science Methodology* (Vol. 5, pp. 23-86). Stamford, CT: JAI Press.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher,, . 31*(3), 24-31.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*: American Psychological Association.

Thurstone, L. (1938). A new rotational method in factor analysis. *Psychometrika, 3*(4), 199-218.

Tucker, L. R. (1951). A method for synthesis of factor analysis studies: EDUCATIONAL TESTING SERVICE PRINCETON NJ.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321-327.

Velicer, W. F., Eaton, C., & Fava, J. (2000). Construct Explication through Factor or Component Analysis: A Review and Evaluation of Alternative Procedures for Determining the Number of Factors or Components. In R. Goffin & E. Helmes (Eds.), *Problems and Solutions in Human Assessment* (pp. 41-71): Springer US.

Voelkl, K. E. (1997). Identification with school. *American Journal of Education*, 294-318.

Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research, 28*(3), 263-311.

Wilkinson, L. (1999). Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist, 54*, 8,594-604.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Wise, S. L. (2006). An Investigation of the Differential Effort Received by Items on a Low-Stakes Computer-Based Test. *Applied Measurement in Education, 19*(2), 95-114.

Wrigley, C., & Neuhaus, J. O. (1955). The matching of two sets of factors. *American psychologist, 10*, 418-419.

Yu, C. H. (2003). Resampling methods: concepts, applications, and justification. *Practical Assessment, Research & Evaluation, 8*(19), 1-23.

Proof

Printed By Createspace

createspace™
an Amazon.com company

Digital Proofer