

Handbook of Quantitative Methods for Detecting Cheating on Tests

Edited by
Gregory J. Cizek and James A. Wollack



Handbook of Quantitative Methods for Detecting Cheating on Tests

The rising reliance on testing in American education and for licensure and certification has been accompanied by an escalation in cheating on tests at all levels. Edited by two of the foremost experts on the subject, the *Handbook of Quantitative Methods for Detecting Cheating on Tests* offers a comprehensive compendium of increasingly sophisticated data forensics used to investigate whether or not cheating has occurred. Written for practitioners, testing professionals, and scholars in testing, measurement, and assessment, this volume builds on the claim that statistical evidence often requires less of an inferential leap to conclude that cheating has taken place than do other, more common, sources of evidence.

This handbook is organized into sections that roughly correspond to the kinds of threats to fair testing represented by different forms of cheating. In Section I, the editors outline the fundamentals and significance of cheating, and they introduce the common datasets to which chapter authors' cheating detection methods were applied. Contributors describe, in Section II, methods for identifying cheating in terms of improbable similarity in test responses, preknowledge and compromised test content, and test tampering. Chapters in Section III concentrate on policy and practical implications of using quantitative detection methods. Synthesis across methodological chapters as well as an overall summary, conclusions, and next steps for the field are the key aspects of the final section.

Gregory J. Cizek is the Guy B. Phillips Distinguished Professor of Educational Measurement and Evaluation in the School of Education at the University of North Carolina, Chapel Hill, USA.

James A. Wollack is Professor of Quantitative Methods in the Educational Psychology Department and Director of Testing and Evaluation Services at the University of Wisconsin, Madison, USA.

Educational Psychology Handbook Series
Series Editor: Patricia A. Alexander

The International Handbook of Collaborative Learning

Edited by Cindy E. Hmelo-Silver, Clark A. Chinn, Carol Chan, and Angela M. O'Donnell

Handbook of Self-Regulation of Learning and Performance

Edited by Barry J. Zimmerman and Dale H. Schunk

Handbook of Research on Learning and Instruction

Edited by Patricia A. Alexander and Richard E. Mayer

Handbook of Motivation at School

Edited by Kathryn Wentzel and Allan Wigfield

International Handbook of Research on Conceptual Change

Edited by Stella Vosniadou

Handbook of Moral and Character Education

Edited by Larry P. Nucci and Darcia Narvaez

Handbook of Positive Psychology in Schools, Second Edition

Edited by Michael Furlong, Rich Gilman, and E. Scott Huebner

Handbook of Emotions in Education

Edited by Reinhard Pekrun and Lisa Linnenbrink-Garcia

Handbook of Moral and Character Education

Edited by Larry Nucci, Tobias Krettenauer, and Darcia Narvaez

International Handbook of Research on Teachers' Beliefs

Edited by Helenrose Fives and Michelle Gregoire Gill

Handbook of Social Influences in School Contexts: Social-Emotional, Motivation, and Cognitive Outcomes

Edited by Kathryn R. Wentzel and Geetha B. Ramani

Handbook of Motivation at School, Second Edition

Edited by Kathryn R. Wentzel and David B. Miele

Handbook of Human and Social Conditions in Assessment

Edited by Gavin T. L. Brown and Lois R. Harris

Handbook of Research on Learning and Instruction, Second Edition

Edited by Patricia Alexander and Richard E. Mayer

Handbook of Quantitative Methods for Detecting Cheating on Tests

Edited by
Gregory J. Cizek and James A. Wollack

First published 2017
by Routledge
711 Third Avenue, New York, NY 10017

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

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

© 2017 Taylor & Francis

The right of Gregory J. Cizek and James A. Wollack to be identified as editors of this work has been asserted by them in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

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

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

Library of Congress Cataloging-in-Publication Data
A catalog record for this book has been requested

ISBN: 978-1-138-82180-4 (hbk)
ISBN: 978-1-138-82181-1 (pbk)
ISBN: 978-1-315-74309-7 (ebk)

Typeset in Minion
by Apex CoVantage, LLC

CONTENTS

Editors' Introduction	ix
Section I INTRODUCTION	1
Chapter 1 Exploring Cheating on Tests: The Context, the Concern, and the Challenges	3
GREGORY J. CIZEK AND JAMES A. WOLLACK	
Section II METHODOLOGIES FOR IDENTIFYING CHEATING ON TESTS	21
Section IIa Detecting Similarity, Answer Copying, and Aberrance	23
Chapter 2 Similarity, Answer Copying, and Aberrance: Understanding the Status Quo	25
CENGIZ ZOPLUOGLU	
Chapter 3 Detecting Potential Collusion Among Individual Examinees Using Similarity Analysis	47
DENNIS D. MAYNES	
Chapter 4 Identifying and Investigating Aberrant Responses Using Psychometrics-Based and Machine Learning-Based Approaches	70
DOYOUNG KIM, ADA WOO, AND PHIL DICKISON	

Section IIb Detecting Preknowledge and Item Compromise	99
Chapter 5 Detecting Preknowledge and Item Compromise: Understanding the Status Quo CAROL A. ECKERLY	101
Chapter 6 Detection of Test Collusion Using Cluster Analysis JAMES A. WOLLACK AND DENNIS D. MAYNES	124
Chapter 7 Detecting Candidate Preknowledge and Compromised Content Using Differential Person and Item Functioning LISA S. O'LEARY AND RUSSELL W. SMITH	151
Chapter 8 Identification of Item Preknowledge by the Methods of Information Theory and Combinatorial Optimization DMITRY BELOV	164
Chapter 9 Using Response Time Data to Detect Compromised Items and/or People KEITH A. BOUGHTON, JESSALYN SMITH, AND HAO REN	177
Section IIc Detecting Unusual Gain Scores and Test Tampering	191
Chapter 10 Detecting Erasures and Unusual Gain Scores: Understanding the Status Quo SCOTT BISHOP AND KARLA EGAN	193
Chapter 11 Detecting Test Tampering at the Group Level JAMES A. WOLLACK AND CAROL A. ECKERLY	214
Chapter 12 A Bayesian Hierarchical Model for Detecting Aberrant Growth at the Group Level WILLIAM P. SKORUPSKI, JOE FITZPATRICK, AND KARLA EGAN	232
Chapter 13 Using Nonlinear Regression to Identify Unusual Performance Level Classification Rates J. MICHAEL CLARK, WILLIAM P. SKORUPSKI, AND STEPHEN MURPHY	245
Chapter 14 Detecting Unexpected Changes in Pass Rates: A Comparison of Two Statistical Approaches MATTHEW GAERTNER AND YUANYUAN (MALENA) McBRIDE	262

Section III	THEORY, PRACTICE, AND THE FUTURE OF QUANTITATIVE DETECTION METHODS	281
Chapter 15	Security Vulnerabilities Facing Next Generation Accountability Testing JOSEPH A. MARTINEAU, DANIEL JURICH, JEFFREY B. HAUGER, AND KRISTEN HUFF	283
Chapter 16	Establishing Baseline Data for Incidents of Misconduct in the NextGen Assessment Environment DEBORAH J. HARRIS AND CHI-YU HUANG	308
Chapter 17	Visual Displays of Test Fraud Data BRETT P. FOLEY	323
Chapter 18	The Case for Bayesian Methods When Investigating Test Fraud WILLIAM P. SKORUPSKI AND HOWARD WAINER	346
Chapter 19	When Numbers Are Not Enough: Collection and Use of Collateral Evidence to Assess the Ethics and Professionalism of Examinees Suspected of Test Fraud MARC J. WEINSTEIN	358
Section IV	CONCLUSIONS	371
Chapter 20	What Have We Learned? LORIN MUELLER, YU ZHANG, AND STEVE FERRARA	373
Chapter 21	The Future of Quantitative Methods for Detecting Cheating: Conclusions, Cautions, and Recommendations JAMES A. WOLLACK AND GREGORY J. CIZEK	390
Appendix A		401
Appendix B: Sample R Code for Data Manipulation and Computing Response Similarity Indices		405
Appendix C: Openbugs Code for Fitting the Bayesian HLM and Estimating Growth Aberrance		415
Contributors		417
Index		423



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

EDITORS' INTRODUCTION

People cheat.

Cheating occurs in the preparation of yearly federal income tax returns to save a few dollars. Newspaper reporters plagiarize to meet a deadline. Professional (and, increasingly, amateur) athletes knowingly use prohibited substances to get a competitive edge. Many of the persons outed in the Ashley Madison database disclosure apparently subscribed to that service to cheat on a spouse.

In numerous pursuits—from stock trading to fantasy football—impermissible inside information is often used to get ahead. For example, as we were writing this introduction, it was reported that diesel-engine Volkswagens had been secretly equipped by the manufacturer with software that could detect when they were undergoing an emissions test and could alter their output so as to pass the tests, and that the emissions, when not in test mode, were up to 40 times higher than allowed.

Surveys of American students at all levels, from elementary school to graduate school, also reveal self-reported cheating on homework assignments and tests. Perhaps the kind of cheating that has received the most recent attention in educational contexts is when educators are found to have changed students' answers on state-mandated achievement tests, either for the purpose of ostensibly helping the students avoid being retained in grade, denied a diploma, or some other consequence, or for the purpose of aiding the educator to obtain a monetary bonus, personal professional recognition, or to help the educators' school meet an overall achievement goal.

It might be argued that some cheating has more serious and far-reaching consequences than other cheating. Indeed, the consequences of a third-grader copying the correct word on a spelling test from a nearby classmate would seem to be of less concern than an aspiring but incompetent physician obtaining an advance copy of the test that will be used to determine whether he or she is licensed to practice medicine. Physician licensure examinations—and other tests with important consequences for individuals, organizations, or societies—are examples of what are often termed *high-stakes tests*. It is these kinds of consequential cheating practices and high-stakes testing situations that are the focus of this book.

Of course, the optimal strategy would always be to prevent cheating that has serious implications from happening in the first place. Both of the editors of this volume (and many other researchers) have written extensively about policies and practices that can

be implemented to prevent cheating on high-stakes contexts. Unfortunately, despite many ounces of prevention, a pound of detection is often necessary.

Perhaps equally unfortunate is the fact that it is not always easy to detect that cheating has occurred. A claim that cheating has occurred typically requires an *inference*—that is, a reasoned conclusion or judgment based on evidence—because cheating is most often a clandestine activity leaving no obvious, direct, objective, and unambiguous signs pointing either to the manner in which the cheating occurred or identifying the person(s) who engaged in the behavior.

Consider what might be the “simplest” of circumstances, in which one test taker is observed to be staring intently at the paper of the classmate seated next to her. After a couple of minutes of sustained peering at her classmate’s paper, she picks up her pencil and immediately bubbles in an answer that subsequent investigation confirms is identical to the one recorded by the classmate. Clearly, this test taker was copying. In this situation, there exists what some might consider to be the best of all possible evidence: the test proctor personally observed the test taker engaging in the behavior. Or did he?

Reasoned conclusions or judgments based on evidence—that is, inferences—can be classified along a scale from low to high. *Low-inference* conclusions are those for which alternative conclusions cannot be supported by any plausible alternative source of evidence, or which are far more likely to be accurate conclusions than any alternative interpretations. *High-inference* conclusions are those for which other sources of evidence could easily be imagined and evidence gathered to refute an initial conclusion, or which can be reasonably challenged by plausible alternative interpretations of the original evidence.

Let us reconsider the ostensibly simple case of a proctor’s direct observation of student copying. The conclusion that the student was even looking at the classmate’s paper requires a substantial inference. For example, the student may have simply been averting or resting her eyes and taking a short break from focusing closely on her test materials. And what was going on in her mind? Was she seeking answers on the classmate’s test or deep in thought about the test question she was working on and contemplating her answer? And, without much more precise observations by the proctor, it cannot be said for certain that the answer the test taker recorded was to the same question where the identical response was noted, or whether she was responding to a completely different test item. In short, a fair amount of inference is necessary to conclude that cheating has occurred in this scenario, and it should be clear that even direct observation is hardly the gold standard criterion for concluding that cheating has taken place, especially when the incidents are isolated.

It is with that important perspective in mind that we approach the use of quantitative methods for detecting cheating. We believe that, in many situations, statistical evidence often requires *less* of an inferential leap to conclude that cheating has occurred than other sources of evidence often putatively and sometimes mistakenly considered to be stronger.

Of course, statistical evidence has limitations and also requires inference to arrive at a conclusion. However, we note that, with the increasing sophistication of quantitative methods for detecting cheating, the inferential leap required to support a conclusion of cheating has been shrinking.

In the past, one of us (GJC) suggested that “it would seem prudent that the weight of statistical evidence only be brought to bear when some other circumstances—that is, a trigger—provide a reason for flagging cases for subsequent statistical analysis” (1999, p. 142). Similarly, since Buss and Novick (1980) first asserted that statistical indices

alone should not be used to invalidate examinees' scores, most articles on cheating detection have included some allusion to the notion that because statistical methods are probabilistic in nature, they ought not to be trusted in isolation. These suggestions may have been reasonable when they were made—when methods to detect cheating on tests were just emerging. Now, however, these suggestions seem outdated, particularly given the methodological advances in cheating detection described in the chapters of this book. These newer methods, comprising a specialization in quantitative analysis to detect cheating called *data forensics*, are now routinely used in the contemporary practice of cheating detection.

To be sure, it would always be desirable to have multiple sources and types of evidence supporting or refuting a concern that cheating has occurred. Cases of answer copying—the context where cheating detection methods first arose—are almost always accompanied by a seating chart or some information in support of the conclusion that two examinees had access to each other's papers. Physical evidence (e.g., answer sheets) is often available in cases of test tampering. Concerns about item preknowledge are often accompanied by some evidence of a common background variable. In other cases, the evidence may come from multiple statistical indexes.

As the methods and practice of quantitative detection of cheating have evolved, it now seems not only entirely defensible but prudent to use statistical analysis as the trigger to prompt additional investigation. And, in some cases, absent compelling counter-evidence, rigorous statistical methods yielding highly improbable findings might alone be sufficient to conclude that an examinee's test score should not be considered a valid representation of his or her knowledge or ability.

What? Statistical evidence alone? Perhaps.

Consider the following situation. Imagine that you are a real estate agent and you are attending the statewide annual convention for your profession. The convention attracts 10,000 real estate agents from around your state for professional development, showcasing new technologies, and personal networking. A much-anticipated feature of the annual meeting each year occurs when slips of paper containing the names of each convention attendee are put into a bucket, and a drawing occurs on the last day of the convention to identify the winner of a new luxury car. The association president for the past several years, Misty Meanor, is officially charged with drawing a single slip of paper from the bucket to identify the winner.

At this year's convention, the president reaches into the bucket, pulls out a slip of paper, and announces the winner: "Misty Meanor." The thousands of real estate agents in attendance for the drawing react, as one might anticipate, with a great deal of skepticism. Statistically speaking, there is a 1 in 10,000 chance that the president would have drawn her own name, so it *is* possible. After some commotion and discussion, it is agreed that the slip of paper will be returned to the bucket and the drawing will be performed again. Again, the president reaches into the bucket, pulls out a slip of paper and announces the winner's name: "Misty Meanor."

Twice in a row? That's pretty unusual. In response, the other officials of the association—and most of the attendees—demand that the drawing be done again. Again, the president reaches into the bucket, pulls out a slip of paper, and announces the winner: "Misty Meanor." Three more times a redrawing of names is demanded. Each time the same result occurs, with the president announcing her own name as winner.

Let us at this point add the perfunctory disclaimers that are, we believe, too often trotted out *de rigueur* in such situations. First, the fact that something *could* happen is frequently misinterpreted to mean that it *did* happen. Despite the infinitesimal

probably of the occurrence, it is true that the president could have randomly drawn her own name six times in a row. That's the thing about random: It can happen. It might snow in Miami in July, maybe once every hundred trillion years. Probabilistically, it *could* happen. But if you were in a coma and woke up in July to see a huge snowstorm outside, you'd have pretty strong reason to conclude that you're not in Miami. Similarly, although statistically it's *possible* that the association president could have pulled out her own name at random, the chances of it happening six times in a row are so remote that there wouldn't be a person at the convention who wouldn't suspect that *some* kind of cheating occurred.

A second obligatory disclaimer is that statistical probability doesn't or can't *prove* that cheating occurred. True—by definition. Statistical methods for detecting cheating are necessarily probabilistic. The methods allow us to quantify the likelihood that something has occurred—or not. What the statistical results *can* do is provide strong evidence that the observed results were not likely to have occurred as a natural result of some normal course of events. The long-held belief within the testing industry that statistical methods ought not be used in isolation for purposes of imposing consequences on examinees dates back to a time when research and practice were just beginning in the emerging field of statistical methods for detecting test cheating. However, the field has now progressed to the point where it is possible to associate statistical decisions with fairly reliable probability statements quantifying the likelihood of our making incorrect conclusions, and in many cases, these probabilities are sufficiently remote that the test scores under review cannot be reasonably considered valid.

Let us return for a moment to our example of the real estate association president repeatedly selecting the slip of paper with her name on it from the 10,000 choices in the bucket. On the one hand, the fact that six times in a row the president was identified as the winner does not *prove* that she cheated. On the other hand, the fact that such an occurrence is so improbable demands some other credible explanation beyond the president's assertion that "it *can* happen." Indeed, there wouldn't be a single conference attendee who wouldn't demand that someone look up the president's sleeve, inspect the names in the hat, independently verify the names that appeared on the selected slips of paper, assign another person to draw the winning name, or pursue any one of myriad other information sources or options to assure a fair result.

A fair result. That is the goal of all high-quality tests. We want those licensed as physicians to be competent to engage in the safe and effective practice of medicine. Those who pass a medical licensure examination should do so because they truly possess the knowledge and skill to perform surgery, not because they attended a test preparation seminar where they had inappropriate prior access to the actual questions that would appear on their licensure examination. We want students to pass a high school exit examination because they have mastered the academic content deemed essential for the conferring of a high school diploma, not because they were seated next to a capable student from whom they were able to copy answers during the test. If educators or school systems are to be compared or judged as more effective based in part on their students' achievement test scores, we want it to be because their students actually learned more, not because an adult erased the students' wrong answers and bubbled in the right ones.

As it turns out, there are quantitative methods for detecting these kinds of behaviors—pretty good ones—and those methods are the focus of this book.

This volume is organized into sections that roughly correspond to the kinds of threats to fair testing represented by different kinds of cheating behaviors. Each section

and subsection begins with a chapter that describes the current state of affairs with respect to strategies for detecting specific types of suspected cheating.

In Section I, we summarize the scope and contexts in which cheating occurs; we proffer a psychometric argument for why cheating should be a significant concern for test makers, test takers, and test score users; and we provide a working definition of cheating grounded in the psychometric concept of validity. In addition, we describe a distinguishing characteristic of this volume—namely, the availability to chapter authors of common datasets to which their cheating detection methods could be applied and results compared across methods. We describe the features of these two datasets: one from a statewide K-12 student achievement testing program, and the other from a national credentialing examination program.

Section II of the book, entitled “Methodologies for Identifying Cheating on Tests,” comprises three groupings of chapters. Section IIa contains chapters that describe methods for detecting improbable similarity in test responses, answer copying, and aberrance. Chapter 2, authored by Cengiz Zopluglu, provides an overview of the status quo in this area. Chapter 3, by Dennis D. Maynes, documents the similarity analysis method, *M4*, used by Caveon Test Security for detecting potential collusion among individual examinees. The focus of Chapter 4, by Doyoung Kim, Ada Woo, and Phil Dickison, is on identifying aberrant responses using both item response and response time methods; however, they also introduce a novel approach from machine learning to identify similar background characteristics among examinees with anomalous response patterns.

Section IIb contains five chapters that describe methods for detecting preknowledge of test content and item compromise. In Chapter 5, Carol A. Eckerly provides an overview of the state of affairs in this area. In Chapter 6, Jim Wollack and Dennis D. Maynes describe a method for detection of test collusion using similarity statistics and cluster analysis. Chapter 7, by Lisa S. O’Leary and Russell W. Smith, provides information on how preknowledge and compromised test content can be detected using differential person and item functioning methods; in Chapter 8, Dmitry Belov shows how preknowledge can be detected using information theory and combinatorial optimization methods. The final chapter in Section IIb, Chapter 9 by Keith A. Boughton, Jessalyn Smith, and Hao Ren, illustrates how response time data can be used to detect compromised items or persons.

The five chapters comprising Section IIc focus on detection of unusual score gains on tests and detecting test tampering. Section IIc opens with an overview by Scott Bishop and Karla Egan in Chapter 10 of the status quo for detecting erasures and unusual score gains. Chapter 11, by Jim Wollack and Carol A. Eckerly, describes a method for detecting group-based test tampering. Billy Skorupski, Joe Fitzpatrick, and Karla Egan describe a Bayesian hierarchical linear model for detecting aberrant growth at the group level in Chapter 12. Chapter 13, by Mike Clark, Billy Skorupski, and Stephen Murphy, details a nonlinear regression approach for identifying unusual changes in performance level classifications. Chapter 14 contains a description and evaluation of two methods for detecting unexpected changes in pass rates by Matt Gaertner and Malena McBride.

Section III of this volume, entitled “Theory, Practice, and the Future of Quantitative Detection Methods,” shifts the focus to policy and practical implications of quantitative detection methods. The first chapter in this section, Chapter 15, by Joseph A. Martineau, Daniel Jurich, Jeffrey B. Hauger, and Kristen Huff, examines security vulnerabilities facing K-12 assessment programs as we enter the next generation of

accountability testing. Chapter 16, by Deborah J. Harris and Chi-Yu Huang, highlights the importance of—and provides suggestions for—establishing baseline data for incidents of misconduct in the next-generation assessment context. Chapter 17, by Brett P. Foley, provides a wealth of information, guidance, and rationale for constructing visual displays for accurately portraying test fraud data. In Chapter 18, Billy Skorupski and Howard Wainer provide an alternative to frequentist methods for detecting cheating in their chapter, which presents the case for Bayesian methods for investigating test fraud. The final chapter in Section III, Chapter 19, by Marc J. Weinstein, provides recommendations for the collection and use of evidence beyond quantitative analysis in the context of concerns about examinee violations of ethical or professionalism norms.

The final section of the book, Section IV, comprises two chapters. In Chapter 20, Lorin Mueller, Yu Zhang, and Steve Ferrara provide a synthesis across methodological chapters and bring together what has been learned from analyses of the common datasets. In the closing chapter of the book, Chapter 21, we provide an overall summary, conclusions, and likely next steps for quantitative methods for detecting test cheating.

We conclude this editors' introduction with some personal observations.

I (JAW) have been studying test security issues since 1995. I was introduced to the subject through my graduate student assistantship in the university testing center. My supervisor told me that a faculty member suspected that one of her students may have been copying answers from a neighboring examinee and asked me to conduct an analysis to test this hypothesis. To my surprise, although there was a significant literature base describing the cheating problem, approaches to detect (or even prevent) security breaches were frighteningly sparse. However, over the course of the next few years, several approaches to detect answer copying were published, followed by the first forays into detecting item preknowledge. By the early 2000s, methods to detect cheating on tests had a regular presence at national conferences, such as the American Educational Research Association and the National Council on Measurement in Education (NCME).

In 2003, the importance of focusing on test security was brought to light by two separate events. The first was the announcement of a new start-up company, Caveon, which would be dedicated entirely to assisting testing programs in preventing and detecting security breaches. Many of the founding fathers from Caveon had spent their careers in the licensure and certification testing industry, which was growing tremendously as a field and had experienced a handful of security breaches. The other was the publication of *Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating* (Jacob & Levitt, 2003). At the time, No Child Left Behind was in its infancy, and although there had been isolated reports of cheating in the public schools, few understood the magnitude of the problem.

In the approximately 10 years since, test security has continued to evolve and expand as an area of concentration. Several organizations (e.g., the Association of Test Publishers [ATP], the Council of State School Officers [CCSSO], and NCME) all released documents focused on best practices to uphold the integrity of test scores in K-12 testing contexts, and the U.S. Department of Education held (and published a report based on) a testing integrity symposium focused on the same topic. ATP held two separate security summits and formed a standing committee to advise members on current and future test security issues, and to provide guidance for how best to protect testing programs against potential security breaches. And research on developing new and improved methodologies to detect different types of cheating in multiple testing

environments has flourished to such a degree that there is now an annual conference focused entirely on advancing test security research and promoting best practices.

Twenty years ago, when I first embarked on this research area, I never imagined that I could make a career out of studying cheating on tests. And, although it has been an exciting area in which to work, it is a bit unsettling that the field has flourished as much as it has. After all, this work would not be possible (or at least not necessary) were it not for a steady and evolving stream of creative, courageous, entrepreneurial, under-prepared, and ethically challenged individuals. Consequently, all these years later, the field is as fertile as ever and appears poised to support active research programs for the foreseeable future.

The publication of this book is the culmination of several years of work and the contributions of many individuals who helped either directly, by providing expertise that is reflected in the volume itself, or indirectly, by assuming other responsibilities that allowed me the time I needed to finish this project. First and foremost, I would like to thank my co-editor, colleague, and friend, Greg Cizek. Greg is probably the single most recognizable and respected individual associated with test security, having—quite literally—written the book on cheating. Working with him on this book has truly been one of my career highlights.

I am indebted to all the authors who contributed to this book and am deeply appreciative not only of the care they took in writing their chapters and responding to editorial comments but in their patience with me for getting access to the common datasets and in awaiting feedback. I also owe a debt of gratitude to the credentialing and education entities who were willing to allow us access to real data from their programs, so that we may enhance our understanding of new and existing methodologies. I understand that preparing these datasets and securing the requisite permissions within your organizations was time-consuming and tricky, and without any direct reward or compensation in return. It is only through your generosity that the benefits of utilizing the common datasets were possible.

My deepest appreciation goes out to our editor at Routledge, Rebecca Novack, who saw fit to work with our idea and with our idiosyncrasies.

I am also very lucky to work at the University of Wisconsin with the best collection of individuals to be found. Special thanks to my colleagues at Testing & Evaluation Services and the Educational Psychology department for endless support, friendship, intellectual curiosity, and collegiality. It's truly a pleasure to come to work each day.

And finally, my family. I am blessed with a loving wife, Jodi, and three amazing daughters, Hannah, Ashley, and Haylie, who inspire me, challenge me, and encourage me in ways I never knew possible.

I (GJC) would like to indicate several points of gratitude. I'll begin with a specific commendation to the folks at Caveon Test Security, Inc. I think that most researchers and practitioners who are familiar with the field of quantitative methods for detecting test cheating will recognize that organization as perhaps the leading provider of examination cheating prevention and detection services. Caveon has recently celebrated 10 years of work in the field of test security and, for perhaps most of those years, I have urged them to provide detailed descriptions of their data forensic methods—professional urging that probably lapsed into not-so-professional nagging at some points.

In the end, it is noteworthy that their widely used approach has been included in this volume. Indeed, each of the chapters in this volume represents the good faith of their authors to describe their methods in such a way as to enable them to be scrutinized in the best traditions of peer review. The use of the common dataset by authors in this

volume allows methods to be compared for the extent to which they identify true case of cheating (statistical power) and guard against identifying a case of cheating when, in fact, cheating did not occur (a false positive). At minimum, testing programs, courts, and test takers should demand that any approach to statistical detection of test cheating used in support of high-stakes, consequential decision making should be transparent, subjected to peer review, and its results should be replicable by an impartial, independent, and qualified entity. It would seem irresponsible to endorse any approach to cheating detection that fails on these basic criteria.

It is appropriate to acknowledge many other debts in the preparation of this book. First, as alluded previously, a volume such as this would not be possible without the professional contributions and academic integrity of so many authors who have been willing to subject their work to the scholarly scrutiny of other researchers and experts in the field. I appreciate the work of each author represented in this volume who is committed to develop, share, and critique quantitative methods used to help ensure the validity of test scores. I am particularly grateful to the enduring scholarly contributions and collegial support of my co-editor, Jim Wollack. If I could ask only one person for advice on a thorny test integrity problem involving quantitative detection methods, it would be Jim. He is one of the most thoughtful, knowledgeable, and experienced researchers in the field of psychometrics generally; in my opinion, he is clearly the leader in this field of data forensics today and will be for the foreseeable future.

In addition, I appreciate the support for this work provided by University of North Carolina at Chapel Hill (UNC). I am deeply grateful to be the recipient of research leave granted by the Dean of the School of Education, Dr. G. Williamson McDiarmid, and by the UNC-Chapel Hill Provost's Office. It is doubtful that such an ambitious project could have been completed without their encouragement and the time allocated to pursue this work. I am also grateful for the professional insights, encouragement, good advice, and friendship of our editor, Rebecca Novak, at Routledge.

Finally, I humbly acknowledge how good God has been to me in providing me with so many other sources of support for this project. I am grateful for the enduring support of my academic family at UNC whom I know as colleagues always willing to provide honest input, creative insights, and reasoned, constructive critique. I am also thankful for the blessings of family, including my parents, Roger and Helen Kabbes, who although they may often be uncertain about what psychometrics is, have never been uncertain about their support for me, and I am deeply grateful for my wife, Julie—a partner, companion, friend, critic, and advocate whose love and encouragement are beyond measure.

GJC/JAW

Chapel Hill, North Carolina and Madison, Wisconsin

REFERENCES

- Buss, W. G., & Novick, M. R. (1980). The detection of cheating on standardized tests: Statistical and legal analysis. *Journal of Law and Education*, 9(1), 1–64.
- Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum.
- Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118, 843–877.

Section I

Introduction



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

1

EXPLORING CHEATING ON TESTS The Context, the Concern, and the Challenges

Gregory J. Cizek and James A. Wollack

Cheating on tests is a pervasive concern across many, varied contexts. Those contexts span a range that includes elementary school achievement testing programs to postsecondary training and professional credentialing programs that offer specialized certifications or licenses to practice in a given field.

It is difficult to obtain firm estimates on the extent of cheating on tests, but by nearly any metric it would seem that the extent of cheating remains high. Biannual and occasional surveys of high school and college students indicate that approximately 60% of high school students in the U.S. admit to cheating on a test at least once per year in 2010, although that figure dropped to 53% in 2012 (Josephson Institute, 2012). In a review of studies involving college students, Whitley (1998) found that an average of 43% of college students reported cheating on exams. Estimates of the extent of cheating on licensure and certification examinations are more difficult to obtain, but because the stakes for such credentialing tests are high, it may be reasonable to expect similar or higher rates of cheating.

Whatever the exact figures, threats to test security are omnipresent. Whenever there are important consequences associated with test performance, there will be constant challenges to test security, and it is likely that any entity offering a credential to candidates who pass a testing requirement will experience attempts to compromise the security of the test.

For example, within the U.S., it has been observed that access to occupations is increasingly licensed and regulated and advancement within occupations is increasingly accompanied by credentialing requirements (Collins, 1979), where credentials play a greater gatekeeping role with regard to entry into and progress within a vocation. A credential refers to a formal “attestation of qualification or competence issued to an individual by a third party (such as an educational institution or an industry or occupational certifying organization) with the relevant authority or assumed competence to issue such a credential” (U.S. Department of Labor, 2014, p. 22666). As the number of occupations that mandate a credential to regulate entrance to and advancement within a profession grows, test security is an increasing concern. According to Wollack

and Fremer, “with tests now serving as the gatekeepers to so many professions, the incentive to cheat is at an all-time high” (2013, p. xi).

In addition to the stakes associated with test performance, the increasing frequency of cheating appears to reflect sociological trends: One public policy researcher has opined that America is becoming a “cheating culture” (Callahan, 2004). However, the prevalence of cheating is not limited to the United States; a regular feature called “Cheating in the News” on the website of a prominent test security company chronicles cheating-related news stories from around the world and across diverse professional and educational contexts (see www.caveon.com/citn/).

Increasing attention to accountability in K-12 education contexts is also apparent. Ushered in most notably by legislation such as the Improving America’s Schools Act (1994) and the No Child Left Behind Act (2001), test performance can have high stakes for students (e.g., graduation, promotion, scholarships) and for teachers (e.g., advancement, salary bonuses, professional recognition), as well as for school districts and states (e.g., making *adequate yearly progress*, closing achievement gaps, resource allocations, and funding).

WHAT IS CHEATING?

There are likely many useful definitions of cheating. In the context of testing, Cizek has defined cheating as “any action taken before, during, or after the administration of a test or assignment, that is intended to gain an unfair advantage or produce inaccurate results” (2012a, p. 3). A few aspects of that definition warrant further elaboration.

First, cheating can occur at nearly any point in the test development, administration, and scoring process. Examinees may attempt to gain inappropriate prior access to test content, even before an examination is administered, either by participating in an unauthorized test preparation course or other activity, by sharing secure test content with other examinees, or by outright attempts to acquire test items via electronic hacking, stealing paper copies of test booklets, or other means. A potential test taker might request inappropriate accommodations for a test that he or she does not truly need or deserve. A test candidate might also arrange for another (ostensibly more able) person to take a test in his or her place, obtaining fraudulent results. An examinee might attempt to gain information from another test taker (e.g., copying, collusion), bring impermissible materials into a testing session, or retrieve information during a scheduled break or some other time during the administration of a test. As these and other illustrations make obvious, the potential opportunities for occasions to cheat are vast.

Second, cheating is purposeful: It is done to obtain a test score that does not represent the examinee’s true level of knowledge or skill, or to gain an unfair advantage in a controlled testing environment. This aspect of the definition highlights that not every instance in which a testing policy or administration regulation is violated constitutes cheating. Two examples illustrate this.

For one example, let us consider rules that might guide test takers. Suppose that a group of examinees were to be seated in a testing room according to a seating plan, and the administration rules prohibited examinees from changing their preassigned seating locations. Further suppose that, contrary to the rules, two examinees changed their seats during the test. One examinee changed a seat to be closer to a friend with whom he had prearranged certain signals for communicating answers to test questions between them. Another examinee changed seats for the purpose of having a better view of a timing clock projected at the front of the room, to move away from another

test taker who was distracting him, to sit in a left-handed desk, or other such reasons. Clearly, the testing rules were broken in both cases. Just as clearly, the first instance was to accomplish cheating; in the second instance—and assuming that the examinee's reason for moving was as stated—the rule was not broken for the purpose of gaining any unfair advantage and would not be labeled as cheating.

For the other example, let us consider rules that might guide test administrators. Suppose that a state administers a high-stakes mathematics achievement test to students in the 6th grade. The state's *Manual for Test Administrators* gives specific directions that teachers must read to their students, with no deviations permitted from the exact language in the script—only verbatim repetition of the specified directions. As students are beginning work, two limited-English-proficient students raise their hands to ask the teacher a question. The first student tells the teacher he does not know the meaning of the word *hypotenuse* in the first question, which asks students to determine a given dimension of a right triangle. In a deviation from the administration rules, the teacher points to the hypotenuse of the triangle shown in the test booklet and reminds the student of the formula for finding the hypotenuse given the length of the other two sides of the triangle. The other limited-English-proficient student raises her hand and asks her teacher what she meant when she said “all of your answers must be *recorded* in the time allotted.” The teacher repeats the directions, but the student still looks confused and worries that she doesn't have anything to make a recording. The teacher—in a deviation from the administration rules—tells the student that, “Oh, you don't need to make a recording. In this case, *recording* an answer just means to write it out or color in the bubble of the answer on your answer sheet.” In both situations, it is clear that the testing rules were broken when the teachers gave the students help that was specifically proscribed in the administration manual. However, whereas it is clear that the first teacher's actions were inappropriate, it is just as clear that the second teacher was not trying to cheat or to artificially inflate the student's test score.

Finally, cheating can be for the benefit of oneself, or for another. It is cheating for a test taker to access impermissible materials during an examination; it is also cheating for a test taker to recall test questions on the day he takes an examination for the purpose of providing inappropriate prior access to test content for a friend who will be taking the same test the next day. It would be cheating for test takers to post recalled questions on the Internet, or for test candidates to enroll in a test preparation course that was known to provide them with harvested test content. We have noted that cheating is an intentional action, but it should also be noted that although cheating is a purposeful behavior for *someone* involved, intent on the part of a person who acquires inappropriate test information isn't required. For example, it is certainly within the realm of possibility that a candidate might take a review course that he or she believes to be legitimate, only to learn later that the course used compromised, live items in its training materials. Or, a candidate's friend who has already tested may pass along information about a test without the candidate having asked for (or wanted) that information.

Why Is Cheating Wrong?

Given the data on the incidence of cheating, concern about the problem is clearly warranted. In the U.S., it appears to be occurring in many contexts, and a lot of people appear to be doing it, including students, proctors, teachers in K-12 school settings, coaching school personnel, and freelancers just trying to make an extra buck. International data suggest that the problem is just as serious—or more so—in countries across

the globe. But, really, is copying an answer or providing some information about test content to a friend really such a bad thing? Is cheating really *wrong*?

As it turns out, cheating can be viewed through different lenses. Many readers of this text might view cheating on tests—as we, the editors do—as one of a set of behaviors to which a moral valence should be ascribed. To be clear, yes: we think cheating is morally wrong; it is appropriate that cheating is negatively stigmatized; and those who engage in intentional cheating behaviors are rightly sanctioned for doing so.

However, rather than expound on the ethical dimensions of cheating—which would take us far afield from our areas of expertise—we will suggest that cheating is wrong (and harmful) from a completely different, technical perspective. That perspective is grounded in the psychometric concept of *validity*. Viewing cheating as a technical concern (and by that, it is not meant to suggest that it is *only* a technical concern) actually liberates attention to cheating from becoming mired in comparative, cultural, contextual, or relativist morality debates. It provides a framework for conceptualizing and addressing the problem within a scientific paradigm that allows for more objective consideration, traditionally accepted standards of evidence, and more broadly acceptable solutions.

CHEATING AS A VALIDITY CONCERN

No other document guides the conduct of testing with greater authority than the guidelines jointly produced by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). That document, the *Standards for Educational and Psychological Testing* (hereafter, *Standards*) is formally endorsed by dozens of associations, agencies, boards, and informally viewed as authoritative by perhaps hundreds of others. There have been seven different editions of the *Standards* over the past 50-plus years, beginning with the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA, 1954) to the current *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). Despite changes in each revision of the *Standards*, common themes are evident, particularly regarding the dependability and accuracy of test-generated data.

One topic—validity—has evolved appreciably (Geisinger, 1992), but the primacy of that characteristic for the science and practice of testing has been consistently endorsed with as close to unanimity as is ever witnessed in the social sciences. In 1961, Ebel referred to validity as “*one of the major deities* in the pantheon of the psychometrist” (p. 640, emphasis added). More than half a century later, the *Standards* proclaim validity to be “*the most fundamental consideration* in developing tests and evaluating tests” (AERA, APA, & NCME, 2014, p. 11, emphasis added). It is hard to overstate the technical importance of something that is so elevated in status and described in such unequivocal terms. If it’s so important, what *is* it?

Simply put, validity refers to the extent to which test scores can be interpreted to mean what they are intended to mean. More technically, validity has been defined as “the degree to which scores on an appropriately administered instrument support inferences about variation in the characteristic that the instrument was developed to measure” (Cizek, 2012b, p. 35). An example may help to illustrate. Suppose that an elementary school student takes a test over the mathematics skills covered by a state’s fifth-grade math curriculum, or a high school junior takes the SAT for college admission, or a candidate takes a computer-based dental board examination to determine his

or her preparation to care for patients. Further suppose that each of these test takers obtained a high score on his or her respective test—sufficiently high to be labeled as “Proficient” in fifth-grade math, “likely to obtain a satisfactory freshman year GPA in college,” or “competent for safe and effective entry-level practice.”

To be sure, we can’t know for certain if any of these examinees is proficient, college ready, or competent but—if a test is well-constructed, administered, and scored—we can be fairly confident in ascribing those labels to their performance. Any test is, after all, only a sample of knowledge, skill, or ability taken at a single point in time. So, conclusions about proficiency, readiness, or competence must necessarily be cautious. In psychometric terms, these conclusions are called *inferences* because an informed judgment about the examinee’s more global status must be made based on the smaller sample of behavior. Good tests are designed to support strong, confident inferences. That is, we want tests that allow us, with great confidence, to make claims about proficiency or competence or whatever characteristic it is that we are trying to measure.

The inferences to be made are suggested by the scores examinees receive on their tests. Low scores suggest inferences of ill-preparation, nonmastery, incompetence in a specified area; higher scores suggest inferences of greater mastery, more skill, increased proficiency, and so on. To the extent there is a body of evidence—including theoretical and empirical evidence—that those inferences are defensible, the test scores are said to have greater validity than scores on tests where the available evidence is weak, absent, or contested. In short, test scores are considered to have validity when the interpretations, conclusions, actions—or *inferences*—we make based on those scores are well-supported by evidence that they are good, correct, or accurate.

A number of features of test development, administration, and scoring contribute to that body of evidence in support of valid test score interpretations. The *Standards* (AERA, APA, & NCME, 2014) mention several sources; Cizek (2012b, 2015) has described others. For example, the fifth-grade mathematics test would only be considered to have strong validity evidence if the questions that appeared on that test were based on the content taught in fifth-grade math classes, if it was listed in the state’s fifth-grade math curriculum, and if it could be found in fifth-grade math textbooks. The *Standards* call this “evidence based on test content” (p. 14). Inferences about college readiness from SAT scores are supported by studies that show higher scores on the SAT are related to higher freshman year GPAs; this kind of validity support is called “evidence based on relations to other variables” (p. 16) in the *Standards*. Inferences about a prospective dentist’s ability to think critically about information supplied by a patient are supported by think-aloud protocols that illuminate the cognitions a candidate uses to conceptualize a problem; this kind of validity support is called “evidence based on response process” (p. 15) by the *Standards*.

Just as there are sources of evidence that can support claims about the validity of test scores, there are also many factors that can weaken the confidence we can have in a given test score. For example, if nonnative speakers of English take the fifth-grade mathematics test and score poorly—not because of their mastery of the fifth-grade math curriculum but because they lack the English language reading ability required to access the test—then scores on that test are not truly valid indicators of their mathematics knowledge. If an SAT test-taker’s performance is less than it might have been because of loud distractions outside the testing center room, then the validity of that score is lessened. If the aspiring dentist’s computer monitor exhibits flickering, making it difficult to clearly see dental pathology on displayed radiographs, then the accuracy of any inference based on that test score is threatened.

Perhaps the connection between validity and cheating on tests is now obvious. There may be disagreement about the ethical dimensions of cheating, but it is uncontestable that cheating represents a threat to the valid interpretation of a test score. When cheating takes place—whether in the form of copying from another test taker, collusion, prior access to secure test materials, inappropriate manipulation of answer documents, or any other form—the resulting test scores are not likely to be an accurate measurement of an examinee's true level of knowledge, skill, or ability. In short, a concern about cheating can be viewed as a psychometric concern about the validity or "interpretive accuracy" of test scores.

Implications of Cheating as a Validity Concern

There are several significant implications that arise when cheating is viewed as subsumed under the broader psychometric umbrella of validity. To begin, those interested in quantitative methods for detecting cheating have routinely been grounded in what might be termed a *legal paradigm*. But statistical investigations are by nature probabilistic, not dispositive. Regrettably, it is too common—and we believe, inappropriate—for statistical investigations to take on a legal perspective when evaluating quantitative evidence that cheating may have occurred.

For example, it is regularly noted that quantitative methods are not able to *prove* that cheating has taken place; they are only able to assign a probability that cheating has occurred. For example, in a sample of papers from a 2013 Conference on Statistical Detection of Potential Test Fraud, authors were careful to note that "large score gains are not a determinative indicator of cheating" (Kao, Zara, Woo, & Chang, 2013, p. 1), "statistical criteria are never 'proof' of cheating" (Skorupski & Egan, 2013, p. 12), and "a statistical procedure . . . can never prove that, for example, teachers in a school are cheating" (Skorupski & Egan, 2013, p. 16).

Of course, those statements are true. However, such statements reflect a probabilistic focus that may be appropriate when, for example, the task at hand is to engage in a legal determination that an event has occurred "beyond a reasonable doubt." For better or worse, specialists in the field of testing rarely have the relevant training or experience to approach cheating from a legal perspective. Nor is a legal perspective necessary or desirable from the standpoint of defensible testing practice. What the field of psychometrics *does* represent is the relevant training and experience to proffer judgments about the trustworthiness of test data. Indeed, psychometrists alone have not merely the technical training and experience but also the professional *obligation* to ensure that the data collected by measurement instrumentation and procedures and reported to relevant audiences (e.g., examinees, institutions, boards and agencies, the public) are *valid*.

Evaluations of a body of evidence—including statistical evidence—for the purpose of judging whether cheating has occurred is squarely *outside* the realm of those with expertise in quantitative methods for detecting potential test fraud. Evaluations of a body of evidence—especially statistical evidence—are squarely within the realm of those charged with developing, administering, and evaluating tests for the extent to which they yield valid scores. In short, the field of psychometrics should adopt a scientific, validity-based approach to test score integrity. For example, who is better qualified that we are to make statements such as these:

"There is strong reason to be concerned that these scores may not be valid."

"It is recommended that additional information be gathered or that testing be replicated before interpreting these scores to mean. . . ."

"The scores of Person A and Person B are of questionable validity and may not represent accurate measurements of their true levels of knowledge and skill. In line with the mission of the organization to protect public health and safety, the scores should not be accepted as valid measurements without further investigation."

Of course, just as evidence is required to arrive at a legal determination, evidence is necessary to support these kinds of psychometric conclusions as well. The sources of evidence to be considered might ordinarily include reports/observations by proctors, test administrators, other test takers, or other persons regarding possible cheating; irregularities or unusual behaviors exhibited by an examinee during test registration or a testing session; unusual similarity in the responses of two or more test takers, beyond what would be predicted by their location, ability, and chance; unusually large increases in an examinee's test performance from one test administration to another, beyond what might be expected due to additional preparation; unusual response behavior by an individual examinee on a test (e.g., unusually long or short response times, unusual number of erasures); incongruence between an examinee's test results and results from other relevant variables (e.g., GPA, attendance, previous test scores); or other relevant information that reduces confidence in the accuracy of an examinee's test score.

It should be noted that the above sources of evidence include a mix of quantitative and qualitative information. As one might imagine with qualitative data sources, it is not possible to specify all the different kinds of testing session irregularities, for example, that might cause a concern about cheating to arise or to specify the degree to which a specific irregularity would need to be present to cause a concern. Likewise, it is not possible—and likely not desirable—to specify the degree to which results from any quantitative analysis should exist for the validity of a test score to be questioned. That is, it is not realistic to establish blanket values for statistical significance, effect size, or other quantitative indicators for detection of test cheating. Sometimes a conclusion that cheating is likely to have occurred may be suggested by a single qualitative or quantitative indicator that is extreme; sometimes that conclusion will be suggested by a historical pattern of indicators; sometimes there will be strong doubt about the validity of a score based on more modest values on a constellation of indicators. The following example is illustrative of the case where a single qualitative indicator was strongly suggestive that scores from a student achievement test lacked validity.

The example is based on results from a state-mandated student achievement testing program administered to all students at a certain grade level. The test was configured to contain both operational (i.e., scored) items that counted toward students' scores as well as pilot-test items that were embedded in the test for the purpose of trying out the items; they did not count toward students' scores. More important, the position of the pilot-test items was not disclosed by the state; that is, neither the test takers (i.e., the students) nor their teachers (who served as the test administrators) were aware of which items counted towards students' scores and which did not. When a concern about the validity of some of the students' test scores arose based on a larger than average number of erasures on students' answer booklets, several statistical tests were conducted.

In addition to the quantitative indicators, visual displays of the data were produced. Inspection of the data provided strong evidence that the students' answer documents may have been tampered with, whereby the students' incorrect responses were erased by another person or persons and changed to correct responses, resulting

in inaccurate and invalid results. Figure 1.1 shows the patterns of erasures on a portion of the test for a group of 25 students. Each row of data in Figure 1.1 represents the answers of a student to a series of test questions. The test consisted of both operational (i.e., scored) and nonoperational (i.e., pilot tested) items. The first 13 items in the left half of the figure were operational items; these items counted toward a student's score and responses to these items are shown in the first 13 columns of the Figure. The figure also shows response for 11 nonoperational items (Items 14–24) included in the test; these items did not count toward students' scores and are shown in the right half of the figure.

Individual entries in the each cell indicated students' responses. A letter in a cell indicates a correct response. For example, the line of responses for the first student listed in Figure 1.1 shows that the student answered "D" to Item 1, which was the correct response. A number entered in a cell indicates that an incorrect answer was given by a student and tells which incorrect option the student chose (1 = A, 2 = B, 3 = C, 4 = D). For example, a "1" is entered as the response of the first student to Item 2, indicating that the student answered that item incorrectly. (He or she chose "A" instead of

Item Numbers – Operational Items													Item Numbers – Pilot Test Items										
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
D	1	3	D	4	B	3	A	C	2	B	2	X	1	D	3	A	1	4	1	2	1	1	A
X	D	3	3	3	4	A	4	1	D	1	C	B	4	2	A	3	4	A	C	2	B	D	4
2	D	3	2	4	1	3	2	1	D	3	2	1	B	1	3	4	C	A	4	2	3	1	2
3	2	B	D	4	B	2	4	1	D	X	X	B	2	4	1	3	D	3	D	4	3	A	3
3	1	1	D	3	4	3	4	1	D	3	C	B	1	3	2	A	1	2	C	4	4	1	A
D	2	B	D	3	B	A	A	C	3	4	C	1	3	4	C	1	D	3	3	1	3	A	3
D	2	B	3	B	B	2	3	1	3	B	C	B	3	2	1	C	D	4	4	X	1	C	B
X	X	1	D	X	B	A	A	1	D	1	C	1	B	2	2	D	C	1	2	A	2	1	3
3	2	B	2	B	3	A	A	1	3	1	C	B	1	2	C	2	2	3	4	2	2	4	2
D	2	4	D	1	3	A	A	1	D	B	X	X	1	2	4	C	D	A	1	D	A	B	C
D	2	B	3	3	1	A	4	2	1	4	C	1	2	4	C	1	1	A	2	4	A	A	4
3	2	1	3	3	1	2	2	1	D	3	C	B	B	2	C	D	1	1	2	A	4	3	3
D	X	X	D	3	B	X	X	X	X	B	C	B	A	A	1	C	1	3	C	D	3	4	B
D	X	B	D	B	B	A	X	1	X	B	C	B	1	4	C	3	D	A	3	2	A	A	A
X	X	B	X	X	B	X	3	C	D	B	C	B	A	A	B	4	D	3	C	D	B	C	B
X	X	B	D	B	A	A	X	X	3	X	B	1	2	1	2	D	A	4	D	A	B	2	2
3	2	B	3	3	B	X	X	X	X	1	C	B	3	4	C	1	D	A	3	C	A	4	3
D	X	B	D	X	B	A	A	C	D	B	C	1	2	4	C	3	1	3	D	4	A	A	4
D	X	B	D	3	B	2	A	X	D	B	C	B	B	2	A	A	4	2	C	C	B	1	2
D	X	B	D	1	B	A	A	1	D	B	C	3	O	3	C	3	D	A	X	4	2	A	3
D	D	B	D	X	X	A	A	C	X	3	C	B	B	C	C	D	C	D	1	A	4	B	3
D	D	B	D	B	X	A	A	X	D	B	C	B	B	2	3	A	C	A	C	C	B	2	A
D	D	B	D	X	B	A	A	C	D	X	C	B	1	3	4	A	4	A	C	C	B	D	A
D	2	B	D	4	B	A	A	C	D	B	C	B	B	C	2	1	C	-	D	A	4	1	B
D	X	X	D	3	B	A	A	C	D	B	C	X	1	2	4	2	2	-	X	-	A	3	1

Figure 1.1 Comparison of answer changes on operational and nonoperational sections of a state-mandated student achievement test

the correct response of that item.) Additional shading and various symbols have been added to the cells in Figure 1.1 to indicate the following:

- X = answer change from wrong to right;
- O = answer change from right to wrong, and
- = answer change from wrong to wrong.

Thus, again considering the responses of the first student shown, we can see that the student changed his or her response on Item 13 from wrong to right.

Visual inspection of Figure 1.1 suggests that wrong-to-right answer changing—indeed, erasures of *any* kind—were far more common on the operational (i.e., scored) items in this test than are erasures on the nonoperational items. Because students are not aware of which questions are operational—that is, which ones count toward their scores and which do not—the results shown in the figure provide an illustration of highly improbable responses, strongly suggesting that typical student response behavior was not occurring and some other factor accounts for these observed findings. Such findings *would* be consistent with the hypothesis of inappropriate alteration of student answer documents. That is, if student answer documents were intentionally altered via wrong-to-right erasures as a means of obtaining higher scores, it would be logical to focus on altering only incorrect responses on items that count toward students' scores. Thus, in this case, a plausible other factor is that of inappropriate adult interference in the testing process, whereby students' wrong answers were erased and correct answers filled in by an adult with access to the test materials.

It should be noted that any credible statistical index would also identify the patterns in Figure 1.1 as exceptionally unlikely, although it is likely that none of the quantitative results would be as persuasive as a single visual inspection of the answer and erasure patterns when communicating the results to a nontechnical audience (see Foley, this volume). Thus, in extreme cases such as the one illustrated in Figure 1.1, it should be a matter of indifference which approach one uses. However, where statistical approaches become invaluable is in those situations that are less extreme. In such situations, visual tests fail because they are too imprecise and nonstandard with respect to placement of critical thresholds, are susceptible to personal bias, and fail to adequately account for critical information such as base rate, item difficulty-by-person ability interactions, and other complexities.

In summary, both qualitative and quantitative methods for detecting test cheating can produce trustworthy and powerful evidence that a test score lacks validity. The many chapters in this volume that focus squarely on quantitative approaches attest to the value of such methods. In the end, however, informed judgments about the possibility of cheating are just that—*informed judgments* that should be based on all available evidence, including both evidence that supports the conclusion that a score may lack validity, and disconfirming evidence, when available, that suggests reasons other than cheating may have resulted in the questioned test score. When the balance of available evidence raises sufficient concern about the validity of a test score, specialists in assessment bear a responsibility to alert consumers of the score of any threats to validity and, when it seems appropriate, to defer or decline certifying that the score should be accepted as accurately representing an examinee's true level of knowledge, skill, or ability.

QUANTITATIVE METHODS FOR PROBING THE VALIDITY OF TEST SCORES

The primary emphasis of this volume is on statistical methods that can be used to investigate the validity of test scores, focusing exclusively on the variety of approaches that can be used to help inform decisions as to whether an examinee's score may have been obtained inappropriately and, hence, lack validity.

In recent years, a number of quantitative methods have been proposed and applied to real test data. Given the consequences to examinees from a finding that cheating has occurred, it would seem essential that any methods used to detect potential cheating be themselves validated and have strong research support. However, only some of the methods have been documented in the psychometric or statistical research literature; others have been openly presented and critiqued in settings such as scholarly conferences and symposia; still others have been developed and applied but have not been subjected to the scrutiny of academic peer review.

Unfortunately, the pressing need to detect cheating—particularly in cases where there may be harmful consequences for public health, safety, or serious ethical concerns—has sometimes meant that detection methods have been developed and implemented before they have been fully vetted by the professional psychometric and statistical community. Important information such as the power of a given method to detect cheating, the false positive rate (i.e., the probability that a method will identify examinees as having cheated when they did not) and the false negative rate (i.e., the probability that a method will fail to identify examinees as having cheated when, in truth, they did) are not known. Furthermore, even among those methods for which the statistical properties, including power, the false positive rate, and the false negative rate, have been examined, comparisons across such studies and generalization to practice are difficult because of differences in simulating conditions, such as the type, amount, and magnitude of cheating, the characteristics of examinees involved in cheating, the test properties (e.g., numbers of items, difficulties of items), the properties of compromised items, and the extent to which various model assumptions are satisfied.

A primary goal of this volume is to provide a forum for the transparent presentation, review and evaluation of quantitative methods for detecting test cheating. To foster that goal, the authors of each chapter presenting new methodologies in this book were asked to present two sources of evidence demonstrating the efficacy of their approaches. The first of these is a simulation study or an application to a dataset of their choosing with known compromise. The purpose of this evidence source is to begin to understand the strengths and vulnerabilities of the methods under settings that are under the careful control of the authors. Simulation studies, in which the compromise status for all people and items is known, provide a unique opportunity to learn about methods' detection rates under a variety of controlled circumstances. However, even though simulation studies are carefully designed to mirror reality, it is often the case that the simulated data fail to capture the complexity and messiness of real data. Consequently, results from simulation studies are often seen as best case scenarios for how methods will behave in practice. The same is typically true of author-selected datasets, many of which were initially identified because the magnitude of cheating is sufficiently extreme that even underpowered methods would identify the majority of anomalies.

Second, authors were asked to apply their methods to one of two real datasets, for purposes of better understanding how their methods would apply in practice

and to allow for comparisons among the approaches presented in this volume. The datasets, described below, comprise real test data from one credentialing testing program and one K-12 student achievement testing program in the U.S. The data sets provide the unique opportunity for chapter authors not only to try out various methods using common datasets but also to compare findings—a key feature that permits users of this volume to identify the peer-reviewed quantitative methods that are best suited for their contexts and to rely on research-based evidence regarding their efficacy.

THE DATASETS

As explained earlier in this chapter, test cheating takes many forms, and methods to detect cheating are equally varied. Similarly, not all programs are equally likely to encounter all types of cheating. Test tampering, for example, in which those involved in the administration of the test alter examinee responses after the exam is finished, has been unearthed in numerous paper-based school accountability testing programs but is virtually unheard of on licensure tests. Test tampering also manifests itself very differently on computer-based assessments. Answer copying remains a fairly common form of cheating on paper-based tests but is less common on computer-based tests, where students see only one item at a time and may be on different items from their neighbors, if their neighbors are even taking the same exam. And copying is less common still in computerized adaptive testing, where neighboring examinees will almost never see the same items at the same time.

In addition, even for those forms of cheating that are commonly encountered by all programs, such as item preknowledge, in which examinees enter the test environment with prior exposure to some of the live item content, programs differ in terms of the data they collect that can be used to look for evidence of misconduct. For example, computer-based testing programs will often collect response time data, whereas paper-based testing programs are limited to item response data. Similarly, the quality of information that can be extracted from repeat examinees may be different in a credentialing environment, where those who pass have no reason to retake an exam, compared to examinees in an educational accountability environment, where all students within a district will retest over multiple years.

As a result, no one dataset can serve the purpose of allowing for all methods to be compared. Therefore, two datasets were made available to authors, with authors asked to use the dataset that was more appropriate to their methodological approach. Because of the sensitive nature of this topic and the potential risk to specific programs were the custodians of these data to become known, a number of steps were taken to mask the identity of the programs. For example, in both cases, an undisclosed proportion of data—both examinees and items—were deleted from the dataset prior to making it available to authors. In the case of the Education dataset, to preserve the nested properties of the data (e.g., districts, schools, students) as much as possible, districts were first randomly sampled for inclusion, followed by schools (within sampled districts), and, finally, students (within sampled schools). Also, the datasets represent a sampling of possible forms that were operational at the same time as the ones analyzed. Finally, all item and person identifiers have been relabeled so that individuals' identities could not be discovered, even if the identity of the program were somehow to become known. Detailed information on the two datasets is provided in the following sections.

The Credentialing Dataset

The credentialing data come from a single year of testing for a computer-based program that tests continuously. The program employs item response theory scoring and uses the Rasch model (1980) to scale and score the test. Two equated forms, generically referred to as Form 1 and Form 2, were provided, each containing 170 scored items. Both operational forms were paired with one of three different 10-item pretest sets, for a total test length of 180 items. The Form 1 scored items (and locations) were identical, regardless of pilot set. Similarly, the operational Form 2 items (and locations) were identical, regardless of pilot set. Forms 1 and 2 shared 87 common items, with each containing 83 scored items that were unique to the form. For the items that were common to the two forms, in all cases, the locations of those items were different on the two forms. Data were available for 1,636 Form 1 candidates and for 1,644 Form 2 candidates.

The credentialing dataset is ideal for studying cheating detection methods because it is known to include examinees who engaged in fraudulent test behavior. At least some of this test fraud was candidates illegally obtaining live test content prior to the exam (e.g., item preknowledge), though other types of misconduct were possible as well. For each form of the exam, there were approximately 50 candidates (46 on Form 1 and 48 on Form 2) who were flagged by the testing company as likely cheaters. Candidates were flagged through a combination of statistical analysis and a careful investigative process that brought in other pieces of information. While all examinees flagged were believed to have engaged in test fraud, it is certainly possible that there were other examinees who ought to have been flagged, but were not.

Similarly, the testing program flagged a total of 64 items because they were believed to be compromised, based on both data forensics and a careful investigation. None of these items were known to be compromised at the time they were administered to candidates.

The compromise status of both items and examinees were known to all authors, so that they could use this information in evaluating the performance of their methods. In addition, the dataset included many other item and examinee variables that could potentially be useful for some of the analyses. For each test form, the dataset included a number of background variables on the candidate, including a unique candidate ID, the number of previous times sitting for the exam, codes corresponding to the country in which the candidate was educated, the state in which he or she applied for licensure, the institution in which he or she received educational training, and the testing center in which the candidate sat for the exam. A set of data were provided relating to the examinee's responses on the test, included information on the form and pretest block attempted by each examinee, item responses to all items, operational and pretest, item correct scores for all items, and response times (in seconds) for all items. In addition, the dataset included background information on the test items, including unique item identifiers, location of the item on both forms, item key, item status (scored vs. pilot), and the number of previous forms in which the item has been used as a scored item. For all scored items, Rasch-based item difficulty parameters were also provided. In addition, background data were provided for the test centers, including the state and country in which the particular site was located. These data could be helpful in case a student tested in a center far from their school, but close to another location where they may know many friends, relatives, and potential co-conspirators (e.g., near the town where he or she grew up).

Except for item-level scores, examinee scores were not provided. Authors were reminded that sufficient data were provided, should they desire to compute raw scores or latent trait estimates.

The K-12 Education Dataset

The education data come from a large state's paper-based fourth- and fifth-grade state math assessments over a two-year period. Authors were provided with data from five equated forms within each year. The state assessments were scaled and scored using the Rasch model (1980). Within each testing year, all forms were horizontally equated to each other and horizontally equated to the grade-level forms from the previous year. However, no vertical scale exists between fourth- and fifth-grade forms; hence, scale scores are interpretable within grade only. Consequently, while this dataset allows for the application of the vast majority of approaches to detect cheating, it is not suitable for methodologies that are sensitive to unusual changes in examinee's test scores over time (e.g., between fourth and fifth grades).

All fourth-grade forms included 59 operational items, and fifth-grade forms included 54 operational items. Students were also administered a number of pilot items; however, only data from operational items were provided. The entire dataset included data for 242,732 unique students. Data for both years were available for 70,168 students. Only Year 1 data were available for 97,782 students, while only Year 2 data were available for 74,782 students. Because most of the methodologies are cross-sectional in nature (looking at one year at a time) rather than longitudinal (attending to changes over time), rather than having each author select the year of his or her choosing, to facilitate comparisons across chapters, authors were asked to make the fifth graders in Year 2 the target sample on which analyses should be focused, bringing in fourth-grade or Year 1 data only as required by the method. There were 2,880 Year 2 students for whom only fifth-grade data were available; hence, the entire dataset of Year 2 fifth graders contained 73,048 students, the vast majority of whom appeared in Year 1 as fourth graders. It is this dataset on which most authors conducted their analyses.

Whereas the credentialing dataset included known compromise at both the item and examinee level, the compromise status of the dataset, from either the item or student/educator perspective is not known. However, the dataset did include some items that had been reused and were at greater risk of being compromised. Of course, item exposure rates bear little on the program's vulnerability to educator tampering. Fortunately, in addition to individual responses to all items, the dataset also included data on erasures. Erasures were captured through the scanning process by looking for "light marks." As such, an erasure was coded provided (a) the intensity of the mark was above a minimal threshold (MIN) and below a maximum threshold (MAX), (b) at least one other item choice was above the MAX threshold, and (c) the difference in intensity between the two marks was above a certain threshold. Taken together, these rules are intended to ensure that (a) the light mark was not merely an inadvertent stray mark but also that it was not the intended answer, (b) the light mark is accompanied by another response that appears to be an intended answer, and (c) the intensity of the two marks can be clearly distinguished.

All variables in the dataset were repeated for each of Years 1 and 2. The dataset included three levels of grouping variables: district, school, and math class, each identified with a unique code, though in the case of the math class code the number was unique only within district/school. It also indicated the students' grade level in each of

the two years, so as to help identify why only one year of data was available for many students (e.g., Year 1 fourth graders who did not enroll for fifth grade in the state's public schools versus Year 1 fifth graders who advanced to 6th grade). For each student, the dataset identified which of the five equated grade-level forms was administered and provided several different summative elements, including the student's scaled score and a Rasch-model-based ability estimate. In addition, each student's proficiency classification (Advanced, Proficient, Partially Proficient, or Not Proficient) was included. In this state from which these data were drawn, students at the Proficient and Advanced levels were considered proficient for purposes of satisfying adequate yearly progress (AYP). In addition, two categorical variables were included to indicate the location of the student's score within the proficiency category (i.e., near the lower cutoff, in the middle of the performance level, or near the upper cutoff) and the magnitude of change from the previous year (i.e., significantly improved, improved, no change, declined, or significantly declined). Tables 1.1 and 1.2 provide more detail on the methods by which the categories were determined.

At the item-level, for each question, the dataset included each student's response along with the answer key as well as the student's original/initial answer. In cases where the original answer was different from the final answer, this was taken as a sign of an erasure and an answer change. Finally, for each item, the dataset included three indicator variables corresponding to whether the student changed the item from a right answer to a wrong answer (RTW), a wrong answer to a right answer (WTR), or from one wrong answer to another (WTW).

Table 1.1 Scaled Score Ranges Within Math Proficiency Categories

Grade	Not Proficient			Partially Proficient		Proficient			Advanced
	Low	Mid	High	Low	High	Low	Mid	High	Mid
4	283–400	401–412	413–422	423–427	428–433	434–443	444–455	456–469	470–539
5	363–482	483–500	501–515	516–522	523–530	531–545	546–563	564–583	584–668

Table 1.2 Change in Math Proficiency Level From Previous Year

		Not Proficient			Partially Proficient		Proficient			Advanced
		Low	Mid	High	Low	High	Low	Mid	High	Mid
		Not Proficient	Low	N	I	I	SI	SI	SI	SI
Partially Proficient	Mid	D	N	I	I	SI	SI	SI	SI	SI
	High	D	D	N	I	I	SI	SI	SI	SI
	Low	SD	D	D	N	I	I	SI	SI	SI
Proficient	High	SD	SD	D	D	N	I	I	SI	SI
	Low	SD	SD	SD	D	D	N	I	I	SI
	Mid	SD	SD	SD	SD	D	D	N	I	I
Advanced	High	SD	SD	SD	SD	SD	D	D	N	I
	Mid	SD	SD	SD	SD	SD	SD	D	D	N

Note: SI = Significantly improved, I = Improved, N = No Change, D = Declined, and SD = Significantly Declined.

Data as a Validation Tool

Although most researchers would agree that for methods to be useful, it is important to demonstrate their utility with real data. Data simulations require that the researcher make a host of assumptions about the way in which the cheating occurs. Examples of common assumptions are that low-ability candidates are more likely to cheat; high-ability candidates are more likely to be sources for answer copiers; difficult items are more likely to be compromised; answer copying is most likely to occur in strings of consecutive items; examinees with preknowledge will answer compromised items correctly with perfect accuracy; all noncheating candidates respond according to an item response model; examinees who copy will always use the answer provided by source; examinees copy from one and only one person; examinees in possession of compromised items will always select the indicated key, if one is provided; and examinees will respond much more quickly to items for which they have preknowledge, among others.

Real data, on the other hand, more accurately capture the complexity of cheating, with respect to both the cheating mechanisms and the magnitude of the compromise. However, the one disadvantage of working with real data is that little, if anything, is known about the status of people or items (e.g., compromised or not), nor about the processes underlying an examinee's responses. Hence, when cheating is undetected, it could reflect a method with low power, but could also indicate that the dataset is largely compromise-free. Similarly, when cheating is detected, it may be because the statistical approach had high power, the dataset had an unusually large effect, or the method yielded an unacceptably high false positive rate.

The common datasets used throughout this book were specifically selected because there is good reason to believe they contain authentic degrees (amounts and magnitudes) of test compromise. However, there is also reason to believe that much remains to be learned about the true compromise status of these data. It is our vision that, through a triangulation of findings from a diverse collection of methods, we will come to learn more about not only the specific methods presented herein but also the datasets themselves. As we increase our understanding about the true status of cheating within these datasets, it is our hope that these data may then be used by others to study the properties of new methodologies. To this end, the chapter by Mueller, Zhang, and Ferrara (this volume) focuses not on specific methods but on the datasets themselves, and what this undertaking has taught us about the true state of cheating hidden in their data strings.

CONCLUSIONS

The problem of cheating on tests is widespread and enduring. When cheating occurs on tests given in elementary and secondary schools, the consumers of those test scores—parents, teachers, policy makers, the public, and the students themselves—get inaccurate information with which to make important decisions about students' strengths and weaknesses, changes to instructional programs, needed remediation, funding, and the effectiveness of educational policies. When cheating occurs on tests given for licensure or certification, candidates who did not cheat can be at a disadvantage, and the public can be at serious risk for harm if unqualified or unsafe candidates obtain the credentials allowing them to practice in a profession. The outcomes are troubling regardless of one's view of the ethical aspects of cheating.

Accordingly, we believe that the most fundamental problem that arises when cheating occurs is that of validity—a cross-cutting concern in the chapters of this volume.

Test scores cannot be interpreted with confidence when their meaning is threatened by the potential that cheating has occurred. As assessment specialists, we take this validity concern seriously. Although it is true that cheating most often affects only a small percentage of test scores, the consequences of inaccurate inferences about examinees' knowledge, skill, or ability can be serious. Indeed, the consequences are so serious—and the detection technologies sufficiently advanced—that we believe attention to detecting cheating is no longer optional for the entities charged with oversight of testing programs: They have the professional responsibility to ensure the integrity of the scores they report by including methods for detecting cheating as one of the standard sources of validity evidence that is routinely collected and analyzed. Further, depending on the nature of the testing program, it may be professionally irresponsible to fail to attend to the possibility that scores were obtained by fraudulent means, given the substantial harm that can accrue to individuals and institutions from cheating.

The practice of testing is an ever-changing technology. The age of the ubiquitous Number 2 pencil, exclusive reliance on multiple-choice question formats, and essay blue books is in the recent, but distant, past. Not only are tests now routinely administered and scored by computer, but test administrations can now take advantage of delivery using a wide variety of devices, in remote locations proctored using advanced technology, incorporating innovative question formats, simulations, and elements of multiplayer gaming, with immediate scoring and reporting using artificial intelligence algorithms.

Despite these advances, it may be disconcerting to contemplate the continuing presence and consequences of cheating. However, two facts provide some consolation. First, although there will always likely be incentives to cheat and persons who succumb to those incentives by inventing new and innovative ways to cheat, the technologies for identifying cheating are also evolving; testing specialists continue to develop more powerful and more accurate methods for detecting inappropriate test taking behaviors. This work helps ensure that scholarships are awarded to the most deserving students, that licenses to practice a profession are awarded only to those with the knowledge and skill deemed necessary for safe practice, and that fundamental fairness in testing is the norm. Second, and perhaps more consoling, is that the examinees who gain an inappropriate advantage or whose test scores were obtained by fraudulent means remain a relatively small proportion of all test takers. In nearly all high-quality testing situations, the majority of those who give or take tests understand the importance of, and have committed themselves to, integrity in testing. The quantitative methods for detecting cheating described in this volume would be largely ineffective if it were otherwise. In the end, we hope that this book serves not only as an aid for detecting those who engage in cheating but also as encouragement to the public regarding the accuracy of test data, and with profound respect for those who take the responsibility of giving and taking tests with integrity.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: Author.
- Callahan, D. (2004). *The cheating culture: Why more Americans are doing wrong to get ahead*. Orlando, FL: Harcourt.

- Cizek, G. J. (2012a, April). *Ensuring the integrity of test scores: Shared responsibilities*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, British Columbia.
- Cizek, G. J. (2012b). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, 17(1), 31–43.
- Cizek, G. J. (2015). Validating test score meaning and defending test score use: Different purposes, different methods. *Assessment in Education*. DOI:10.1080/0969594X.2015.1063479
- Collins, R. (1979). *The credential society: An historical sociology of education and stratification*. New York: Academic.
- Ebel, R. L. (1961). Must all tests be valid? *American Psychologist*, 16, 640–647.
- Geisinger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologist*, 27, 197–222.
- Improving America's Schools Act of 1994. Public Law No. 103-382 (20 U.S.C. 6301).
- Josephson Institute. (2012). *2012 report card on the ethics of American youth*. Los Angeles, CA: Author.
- Kao, S., Zara, T., Woo, A., & Chang, C. (2013, October). *Analysis of ability changes for repeating examinees using latent growth curve models*. Paper presented at the 2nd Annual Conference on Statistical Detection of Test Fraud, Madison, WI.
- No Child Left Behind [NCLB] Act of 2001. Public Law No. 107-110 (20 U.S.C. 6311).
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*, with foreword and afterword by B. D. Wright. Chicago: University of Chicago Press.
- Skorupski, W. P., & Egan, K. (2013, October). *A Bayesian hierarchical linear model for detecting group-level cheating and aberrance*. Paper presented at the 2nd Annual Conference on Statistical Detection of Test Fraud, Madison, WI.
- United States Department of Labor. (2014). Request for information on adoption of career pathways approaches for the delivery of education, training, employment, and human services. *Federal Register*, 79(78), 22662–22667. Washington, DC: U.S. Government Printing Office.
- Whitley, B. E. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education*, 39(3), 235–274.
- Wollack, J. A., & Fremer, J. J. (2013). *Handbook of test security*. New York: Routledge.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Section II

Methodologies for Identifying Cheating on Tests



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Section IIa

Detecting Similarity, Answer Copying, and Aberrance



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

2

SIMILARITY, ANSWER COPYING, AND ABERRANCE

Understanding the Status Quo

Cengiz Zopluoglu



INTRODUCTION

In an era of high-stakes testing, maintaining the integrity of test scores has become an important issue and another aspect of validity. A search on the Web with words “test fraud” and “cheating” reveals increasing numbers of news stories in the local and national media outlets, potentially leading to less public confidence about the use of test scores for high-stakes decisions. To address the increasing concerns about the integrity of test scores, the scholarly community is beginning to develop a variety of best practices for preventing, detecting, and investigating testing irregularities. For instance, the National Council on Measurement in Education (2012) released a handbook on test and data integrity, the Council of Chief State School Officials published a test security guidebook (Olson & Fremer, 2013), and the Association of Test Publishers and the National College Testing Association have recently developed a best practices document related to test proctoring (ATP & NCTA, 2015). A symposium on test integrity was hosted with a number of experts from universities, testing companies, state educational agencies, law firms, and nonprofit organizations (U.S. Department of Education, IES, & NCES, 2013). Similarly, an annual scholarly conference on statistical detection of test fraud has been held since 2012 with a growing national and international attention. All these efforts provided an environment for people to discuss the best practices and policies to prevent, detect, and investigate testing irregularities and to ensure the integrity of test scores.

Unusual response similarity among test takers or aberrant response patterns are types of irregularities which may occur in testing data and be indicators of potential test fraud (e.g., examinees copy responses from other examinees, send text messages or prearranged signals among themselves for the correct response). Although a number of survey studies already support the fact that copying/sharing responses among students is very common at different levels of education (e.g., Bopp, Gleason, & Misicka, 2001; Brimble & Clarke, 2005; Hughes & McCabe, 2006; Jensen, Arnett, Feldman, & Cauffman, 2002; Lin & Wen, 2007; McCabe, 2001; Rakovski & Levy, 2007; Vandehey,

Diekhoff, & LaBeff, 2007; Whitley, 1998), one striking statistic comes from a biannual survey administered by the Josephson Institute of Ethics in 2006, 2008, 2010, and 2012 to more than 20,000 middle and high school students. A particular question in these surveys was how many times students cheated on a test in the past year, and more than 50% of the students reported they had cheated at least once whereas about 30% to 35% of the students reported they had cheated two or more times on tests in all these years. The latest cheating scandals in schools and the research literature on the frequency of answer copying behavior at different levels of education reinforce the fact that comprehensive data forensics analysis is not a choice, but a necessity for state and local educational agencies.

Although data forensics analysis has recently been a hot topic in the field of educational measurement, scholars have developed interest in detecting potential frauds on tests as early as the 1920s (Bird, 1927, 1929), just after multiple-choice tests started being used in academic settings (Gregory, 2004). Since the 1920s, the literature on statistical methods to identify unusual response similarity or aberrant response patterns has expanded immensely and evolved from very simple ideas to more sophisticated modeling of item response data. The rest of the chapter will first provide a historical and technical overview of these methods proposed to detect unusual response similarity and aberrant response patterns, then describe a simulation study investigating the performance of some of these methods under both nominal and dichotomous response outcomes, and finally demonstrate the potential use of these methods in the real common datasets provided for the current book.

A REVIEW OF THE STATUS QUO

As shown in Table 2.1, the literature on statistical methods of detecting answer copying/sharing can be examined in two main categories: response similarity indices and person-fit indices. Whereas the response similarity indices analyze the degree of agreement between two response vectors, person-fit indices examine whether or not a single response vector is aligned with a certain response model. Response similarity indices can be further classified based on two attributes: (a) the reference statistical distribution they rely on and (b) evidence of answer copying being used when computing the likelihood of agreement between two response vectors. The current section will briefly describe and give an overview for some of these indices.

Person-Fit Indices

The idea of using person-fit indices in detecting answer copying has been present for a quite long time (e.g., Levine & Rubin, 1979); however, it has not received as much attention as the response similarity indices in the literature with respect to detection of answer copying. The use and effectiveness of person-fit indices in detecting answer copying is a relatively underresearched area compared to the response similarity indices. This is likely because already existing studies had found person-fit indices underpowered specifically in detecting answer copying, a finding that may discourage from further research. One reason of underpowering is probably the fact that most copiers have aberrant response patterns, but not all examinees with aberrant response patterns are copiers. Aberrant response patterns may occur based on many different reasons, and therefore it is very difficult to trigger a fraud claim without demonstrating an

Table 2.1 Overview of Statistical Methods Proposed for Detecting Answer Copying

Response Similarity Indices	
Evidence of Answer Copying	
Statistical Distribution	Number of Identical Incorrect Responses
Normal Distribution	Number of Identical Correct and Incorrect Responses Wesolowsky (2000) ω (Wollack, 1997)
Binomial Distribution	IC (Anikeef, 1954) K (Kling, 1979, cited in Saretzky, 1984) ESA (Bellezza & Bellezza, 1989) K_1 and K_2 (Sotaridona & Meijer, 2002) S_1 (Sotaridona & Meijer, 2003)
Poisson Distribution	S_2 (Sotaridona & Meijer, 2003) GBT (van der Linden & Sotaridona, 2006) B_M (Bay, 1995)
Compound Binomial Distribution	P (Cody, 1985) CP (Hanson, Harris, & Brennan, 1987)
Empirical Null Distribution	t_E (Dickinson, 1945) B and H (Angoff, 1972) Pair I and Pair II (Hanson, Harris, & Brennan, 1987) VM index (Belov, 2011)

*See Karabatsos (2003)

for more in this category.

H^T (Sijtsma & Meijer, 1992)
 D (Trabin & Weiss, 1983)

C (Sato, 1975)

MCI (Harnisch & Linn, 1981)

U3 (van der Flier, 1980)
 Iz (Drasgow et al., 1985)

g_2 (Frary, Tideman, & Watts, 1977)

ω (Wollack, 1997)

C (Sato, 1975)

U3 (van der Flier, 1980)

Iz (Drasgow et al., 1985)

$t_W * tR$ (Sauper, 1960)

B_M (Bay, 1995)

unusual response similarity with another examinee within some physical proximity (Wollack, 1997).

With this practical challenge of using person-fit indices for detecting answer copying in mind, a few studies examined how such indices would perform in detecting copiers under certain conditions (de la Torre & Deng, 2008; Dimitrov & Smith, 2006; Karabatsos, 2003; Meijer, Molenaar, & Sijtsma, 1994; Wollack, 1997). Among these studies, a very comprehensive study by Karabatsos (2003) compared the performance of 36 different person-fit statistics in detecting answer copying. The results indicated that only two indices, H^T (Sijtsma, 1986; Sijtsma & Meijer, 1992) and D_θ (Trabin & Weiss, 1983), showed acceptable performance as measured by the area under the Receiver Operating Characteristic (ROC) curve in detecting aberrant response vectors contaminated with answer copying. The H^T statistic measures how much a person's response pattern deviates from the response vectors of the remaining sample of examinees, and is computed for an examinee i as

$$H^T = \frac{\sum_{i \neq j} \sigma_{ij}}{\sum_{i \neq j} \min\{\beta_i(1 - \beta_j), (1 - \beta_j)\beta_i\}},$$

where σ_{ij} is the covariance between item scores of examinees i and j , β_i and β_j are the proportion of items with correct responses for examinee i and j . The H^T statistic ranges from 0 to 1, and a cut-off point of .3 was recommended by Sijtsma and Meijer (1992) to identify aberrant response vectors (e.g., $H^T < .3$) while Karabatsos (2003) reported that a cut-off point of .22 optimized its classification accuracy. Relatively better performance of the H^T statistic over other indices was also reported by Dimitrov and Smith (2006). The D_θ statistic is simply the average squared residual between an examinee's observed item score and its model predicted probability. Karabatsos (2003) found that a cut-off point of .55 optimized the classification accuracy of aberrant response vectors, (e.g., $D_\theta > .55$).

Another commonly used person-fit statistic for detecting answer copying is the standardized log-likelihood of a response vector (l_z). Although l_z was found performing poorly in some studies (Karabatsos, 2003; Wollack, 1997) for detecting copiers, there were some promising results in another study by de la Torre and Deng (2008). The l_z statistic is computed as

$$l_z = \frac{l_0 - E(l_0)}{\sigma(l_0)},$$

where l_0 is the observed log-likelihood of the response vector, and $E(l_0)$ and $\sigma(l_0)$ are its expected value and standard error. These elements are equal to

$$l_0 = l(X | \theta) = \sum_{k=1}^K (X_k \ln P_k(\theta) + (1 - X_k) \ln(1 - P_k(\theta))),$$

$$E(l_0) = l(X | \theta) = \sum_{k=1}^K (P_k(\theta) \ln P_k(\theta) + (1 - P_k(\theta)) \ln(1 - P_k(\theta))), \text{ and}$$

$$\sigma^2(l_0) = l(X | \theta) = \sum_{k=1}^K P_k(\theta)(1 - P_k(\theta)) \left(\ln \frac{P_k(\theta)}{1 - P_k(\theta)} \right)^2,$$

respectively, where X_k is the observed response for the k th item for an examinee and $P_k(\theta)$ is the probability of correct response for the k th item for the examinee with ability θ . The true ability θ is typically replaced by its estimate ($\hat{\theta}$) in these calculations. It should be noted that de la Torre and Deng (2008) proposed a modified procedure by correcting $\hat{\theta}$ for unreliability and adjusting the corresponding reference distribution by constructing an empirical null distribution through simulation. Although this modified procedure seems to be computationally laborious, results indicated better control of type I error rates and improved power when modified l_z (l_z^*) was used in detecting copiers.

Response Similarity Indices

Rather than looking if a person's response vector is aligned with a response model, response similarity indices focus on the likelihood of agreement between two response vectors under the assumption of independent responding, and these indices differ in how they utilize the evidence of agreement. While some indices only use matching incorrect responses between two response vectors, some use both matching correct and incorrect responses as evidence of agreement. On the other hand, some other indices take all items into account and use matching responses as evidence of copying while using nonmatching responses as evidence of no copying.¹ The indices also differ in their assumptions and reference statistical distributions used for computing the likelihood of observed agreement. While some indices rely on developing and utilizing empirical null distributions, others use different distributional forms such as normal, binomial, Poisson, or compound binomial as a reference when computing the likelihood of observed agreement.

Indices Based on Empirical Null Distributions

The process of generating an empirical null distribution for the number of identical correct/incorrect responses typically starts with matching two examinees from different classrooms, test centers, or geographical locations for the same test, so that pairs of examinees for which copying between was not possible are obtained. Then, the empirical distribution of the number of identical correct and/or incorrect responses between pairs of examinees is developed and treated as a null distribution for future operational use. When a pair of examinees is suspected of answer copying, observed number of matches for the suspected pair is compared to the corresponding null distribution, and empirical probability of observing m or more matches is computed as the upper tail of the null distribution.

The idea of empirical null distributions was first used by Bird (1927, 1929), who constructed his empirical null distribution for the number of identical errors by randomly pairing examinees from different geographical locations and counting the number of identical errors for each pair. Bird used the average of this distribution as a reference. A pair of examinees with an observed number of identical errors larger than this average was flagged as an unusual degree of agreement. Dickenson (1945) similarly derived a formula of "identical error percentage" from his observations based on empirical data. Dickenson (1945) suggested a simple formula $IE = \frac{C-1}{C^2}$, where C is the number of alternatives for the multiple-choice items in the test, as a chance expectation of identical errors between two examinees. He recommended that an observed number

of identical errors larger than 2^*IE be taken as an indication of unusual agreement. For instance, if the items in the test have four alternatives, the chance expectation for the identical error percentage would be equal to 18.7% leading an upper boundary of 37.5%. Therefore, in a 40-item test, any pair of examinees sharing 15 or more identical errors would be deemed suspicious.

Following these early attempts, Saupe (1960) suggested a more statistically sound approach using regression. Suppose that W_{ij} denotes the observed number of items both the i th and j th examinees answer incorrectly and w_{ij} denotes the observed number of identical incorrect responses observed for these two examinees. Similarly, let R_i and R_j denote the observed number of correct responses for the i th and j th examinees and R_{ij} denote the number of identical correct responses between two examinees. For a number of honest pairs for which answer copying did not occur, Saupe (1960) proposed the following regression equations to predict the expected number of identical correct or incorrect responses for any two examinees conditioning on W_{ij} and $R_i * R_j$:

$$\hat{w}_{ij} = \beta_0 + \beta_1 W_{ij}$$

$$\hat{R}_{ij} = \beta_0 + \beta_1 (R_i * R_j).$$

Saupe (1960) proposed to compare observed number of identical correct and incorrect responses between two examinees to their expected values using a t -test:

$$t_w = \frac{w_{ij} - \hat{w}_{ij}}{\sigma_\varepsilon} \quad \text{and} \quad t_r = \frac{R_{ij} - \hat{R}_{ij}}{\sigma_\varepsilon},$$

where σ_ε is the standard error of estimate from the corresponding regression analysis. Associated p values can be obtained from a t distribution with degrees of freedom $N - 2$, where N is the number of honest pairs used in the regression analysis. Saupe (1960) argued that these two tests are independent of each other, and therefore the p -values from two tests may be multiplied to test the hypothesis of no copying using a certain alpha level.

Another approach in this category was proposed by Angoff (1972, 1974), who developed eight different indices similar to Saupe (1960). Based on the empirical investigation of known and admitted copiers and practical considerations, Angoff (1972) suggested two of these indices, B and H , as they were among the most powerful in identifying the known and admitted cheating and were not overly correlated with each other. Angoff's B index uses w_{ij} as the dependent variable and $W_i * W_j$ as the independent variable, where W_i and W_j are number-incorrect scores for the i th and j th examinees. Instead of using a regression approach, Angoff (1972) proposed to compute $W_i * W_j$ for a number of honest pairs, to divide the range of $W_i * W_j$ into equal intervals, and then compute the conditional mean and standard deviation of w_{ij} for the honest pairs within each interval. When there is a suspected pair of examinees, $W_i * W_j$ for the pair is computed to determine to which interval group they belong, and then the observed w_{ij} between two examinees is compared to the mean and standard deviation of w_{ij} for that particular interval to obtain a t statistic. This is analogous to running a regression analysis as what Saupe (1960) did with an independent variable $W_i * W_j$ instead of W_{ij} . The procedure for Angoff's H is similar to Angoff's B , but the dependent and independent variables of interest are different because Angoff's H targets a particular type of copying, which has been labeled as "string copying" in the literature, in which an examinee copies many items in succession. The dependent variable of interest is the

longest run of identically marked incorrect responses and omits between two examinees, and the independent variable is the number of incorrect and omitted responses of the suspected source examinee.

Although the idea of using empirical null distributions is very attractive, it requires access to a large dataset and information on some other variable (e.g., test site) that can be used to pair examinees in such a way that copying between those examinees was not possible, so researchers confidently assert that it's a null distribution. These indices are relatively underresearched in the literature, and this is most likely because researchers do not have access to enough empirical data to create adequate empirical null distributions. Also, it is very likely that these empirical null distributions are test specific, and they cannot be used interchangeably across test forms with different subject areas, different administration times, or different sets of items. Therefore, empirical null distributions have to be created for each specific test before any operational use.

K Index and Its Variants

K index was originally developed by Frederick Kling. There is no publication for the original development of the *K* index; however, Holland (1996) eventually published a study on the theoretical underpinnings of the *K* index. A few years later, Sotaridona and Meijer (2002, 2003) extended this work, developing the K_1 , K_2 , S_1 , and S_2 indices. K , K_1 , and K_2 use the binomial distribution to compute the likelihood of matching on w_{ij} or more identical incorrect responses between two response vectors:

$$\sum_{u=w_{ij}}^{W_s} C \binom{W_s}{u} P_r^u (1-P_r)^{W_s-u},$$

where W_s is the number-incorrect score for the suspected source examinee, $C \binom{W_s}{u}$ is the number of all possible combinations for u matches on W_s -incorrect responses, and P_r is the binomial probability of matching on an identical incorrect response with the suspected source examinee for examinees in the r th incorrect-score group, where r is the number-incorrect score of the suspected copier examinee. The K , K_1 , and K_2 indices differ in the way they estimate P_r . The K index estimates P_r as the average of the number of identical incorrect responses between each examinee in the r th number-incorrect score group and the suspected source examinee. K_1 and K_2 use linear and quadratic regression equations, respectively, to predict P_r from all number-incorrect score groups, rather than using information only from the number-incorrect score group in which the suspected copier examinee belongs. It should be noted that two other indices, Error Similarity Analysis (ESA; Bellezza & Bellezza, 1989) and Index of Collaboration (IC; Anikeef, 1954), are somewhat antecedent to K , K_1 , and K_2 because they also use a binomial distribution to model the number of identical incorrect responses between two examinees; however, the methods of estimating P_r for the ESA and IC indices are less sophisticated and overly simplistic, each making a number of unrealistic assumptions.

Instead of using a linear or quadratic regression, the S_1 index uses a log-linear model to predict the number of identical incorrect responses between the suspected copier and source examinees, and then uses a Poisson distribution to compute the likelihood of matching on w_{ij} or more identical incorrect responses between two response vectors:

$$S_1 = \sum_{u=w_{ij}}^{W_s} \frac{e^{-\hat{M}_r} (\hat{M}_r)^u}{u!},$$

where \hat{M}_r is the model predicted number of identical incorrect responses with the suspected source examinee for the r th incorrect-score group. S_2 is similar to the S_1 index, but with one exception. Instead of using the predicted average number of identical incorrect responses (\hat{M}_r), S_2 uses both the predicted average number of identical incorrect responses and predicted average weighted number of identical correct responses between two response vectors:

$$S_2 = \sum_{u=w_{ij}^*}^K \frac{e^{-\hat{M}_r^*} (\hat{M}_r^*)^u}{u!},$$

where w_{ij}^* is the sum of the observed number of identical incorrect and weighted identical correct responses between two examinees, and \hat{M}_r^* is the model predicted sum of identical incorrect and weighted identical correct responses for the r th incorrect score group.

Generalized Binomial Test and ω Index

The Generalized Binomial Test (GBT; van der Linden & Sotaridona, 2006) uses the compound binomial distribution for the number of identical correct and incorrect responses between two examinees. Let P_{M_k} be the probability of matching for the i th and j th examinees on the k th item, and computed as

$$P_{M_k} = \sum_{o=1}^O P_{iko} * P_{jko},$$

where P_{iko} and P_{jko} are the probability of selecting the o th response alternative of the k th item for the i th and j th examinees, respectively. Then, the probability of observing exactly n matches on K items between two response vectors is equal to

$$f_K(n) = \sum \left(\prod_{k=1}^K P_{M_k}^{u_k} (1 - P_{M_k})^{1-u_k} \right),$$

where u_k is equal to 1 if two examinees have identical responses to item k , and zero otherwise; and the summation is over all possible combinations of n matches on K items. The GBT index is the upper tail of the compound binomial distribution, and the probability of observing $w_{ij} + R_{ij}$ or more matches on K items is equal to

$$\sum_{n=w_{ij}+R_{ij}}^K f_K(n).$$

It should be noted that Bay (1995) proposed a similar idea of using the compound binomial distribution; however, the choice of P_{iko} and P_{jko} was rudimentary, while GBT is estimating these probabilities from a known response model (e.g., Nominal IRT Model).

The ω index (Wollack, 1997) is a normal approximation to the compound binomial distribution for the number of identical correct and incorrect responses between two response vectors. One important difference to note is that ω is a copying statistic (asymmetric), whereas GBT is a similarity statistic (symmetric): ω can be used for similarity by computing in two directions (Test Taker X is copying from Test

Taker Y, and Test Taker Y is copying from Test Taker X) and splitting the alpha level, but it was designed as a directional copying detection index. For the ω index, the observed agreement between two response vectors is compared to its expectation. The expected agreement between two examinees is computed as the sum of probabilities for the suspected copier examinee to give the suspected source examinee's responses, and equal to

$$E_{ij} = \sum_{k=1}^K P_{iko},$$

where o is assumed to be the response alternative chosen by the source examinee for the k th item. The variance of this expectation is equal to

$$\sigma^2 = \sum_{k=1}^K P_{ik} * (1 - P_{ik}),$$

and the ω index is computed as

$$\omega = \frac{(w_{ij} + R_{ij}) - E_{ij}}{\sigma},$$

and compared to the standard normal distribution. The g_2 index (Frary, Tideman, & Watts, 1977) and the procedure proposed by Wesolowsky (2000) are also similar to the ω index, but they differ in estimating P_{ik} .

Summary of Relevant Research Findings

Given the large number of analytical methods proposed in the literature, the empirical type I error rate and power are two critical components to consider when these methods are used in practice. In this context, type I error rate is the proportion of honest pairs falsely detected as copiers (false positives), and power is the proportion of copying pairs truly detected (true positives) by an analytical method in the long run. A large body of research literature exists on how useful these methods are in practical settings, and most of these studies use Monte Carlo simulation as a technique because it provides the most flexible environment to investigate the performance of these indices under many different conditions (Bay, 1995; Belov, 2011; de la Torre & Deng, 2008; Dimitrov & Smith, 2006; Hanson, Harris, & Brennan, 1987; Karabatsos, 2003; Sotaridona & Meijer, 2002, 2003; Wollack, 1997, 2003, 2006; Wollack & Cohen, 1998; Zopluoglu & Davenport, 2012). Among response similarity indices, the ω and GBT indices utilizing item response theory (IRT) statistics, K index and its variants as their non-IRT counterparts, and the VM index are found to be performing well in terms of statistical power and holding the type I error rates at or below the nominal level. Although person-fit indices have generally been found to have limited use in detecting copying, the H^T and the D_0 statistics appear to perform well enough to be useful in some contexts. It should also be noted that Belov and Armstrong (2010) recently suggested a two-stage approach that first screens the sample of examinees using a person-fit index to identify potential copiers and then applies a response similarity index to check whether or not there is unusual agreement between the potential copiers and examinees within proximity.

A REAL-DATA SIMULATION STUDY

A Monte Carlo study was designed to demonstrate the performance of methods briefly summarized in the previous section. The data come from a single year of testing for a computer-based licensure program provided to the authors. The original dataset had two forms, with 87 common items. Also, 94 examinees had been flagged as potentially engaged in test fraud. To demonstrate the performance of person-fit and response similarity indices in a real-data simulation study, the datasets from the two forms were merged, excluding the 94 flagged examinees and dropping any items that were not common to both forms. As a result, the final dataset used for this simulation contained a total of 3,186 examinees' responses to 87 items. Initial exploration of these 87 common items revealed that the item difficulty ranged from .23 to .96 with a mean of .73, and item point-biserial correlations ranged from .05 to .43 with a mean of .24. There were 12 items with low point-biserial correlations (all lower than .15); therefore, these items were removed from the merged dataset, leaving 75 items for further analysis. After removal of the 12 items with low point-biserial correlations, the item difficulty ranged from .23 to .96 with a mean of .74, and item point-biserial correlations ranged from .15 to .45 with a mean of .26 for the remaining 75 items. The Nominal Response Model (NRM; Bock, 1972) was fitted to nominal response data, and dichotomous IRT models (1-PL, 2-PL, and 3-PL; Birnbaum, 1968) were fitted to scored responses for the 75 items. The NRM was fitted using Multilog (Thissen, Chen, & Bock, 2003) with default options, and the maximum number of E-steps was increased to 10,000 to ensure convergence. The NRM item parameters are presented in Appendix A (see Table A.1). The dichotomous IRT models were fitted to the dichotomous dataset using IRTPRO (Cai, du Toit, & Thissen, 2011) with default options while the maximum number of E-steps was increased to 10,000 to ensure convergence, and a prior beta distribution with parameters 5 and 17 was specified for the guessing parameter when fitting the 3-PL model. Among the dichotomous IRT models, the 2-PL model provided the best fit based on AIC and BIC, as shown in Table 2.2. The 2-PL item parameters are presented in Appendix A (see Table A.2). For further analysis, NRM was used for nominal data and 2-PL model was used for dichotomous data when computing IRT-based indices (e.g., ω and GBT).

Simulation Design

The common dataset described above with both nominal and dichotomous responses was used in the study. There were three independent variables manipulated in the study for both the nominal and dichotomous datasets: amount of copying (20%, 40%, and 60%), ability of copier examinee (low vs. high), and ability of source examinee

Table 2.2 Summary Statistics From Fitting Dichotomous IRT Models to Real Dataset

	a		b		c			AIC	BIC
	Mean	SD	Mean	SD	Mean	SD			
1-PL	0.58	-	-2.18	1.56	-	-	238,607.5	239,068.5	
2-PL	0.63	0.27	-2.09	1.52	-	-	237,080.3	237,990.3	
3-PL	0.74	0.31	-1.33	1.44	.21	.06	237,209.3	238,574.3	

(low vs. high). For this 75-item test, a copier examinee copied 15, 30, or 45 items from a source examinee for 20%, 40%, or 60% copying conditions. An examinee with an estimated ability level below 0 was considered as a low-ability examinee, and an examinee with an estimated ability level above 0 was considered as a high-ability examinee. Note that the flagged examinees were eliminated from the dataset to remove as much known aberrance as possible, all remaining examinees in the dataset are assumed to have not copied, and all theta estimates for the remaining examinees are assumed to be unbiased. All independent variables were fully crossed, yielding a total of 12 conditions for each of the dichotomous and nominal response datasets.

To simulate a realistic scenario, each replication contained a number of copying pairs with different ability-level combinations and different amounts of copying. The following process was executed for one replication for both the nominal and dichotomous datasets independently:

1. Generating answer copying pairs: An examinee with low/high ability was randomly selected and then matched with another randomly selected low/high ability examinee from a different test center. A number of items (20%, 40%, or 60%) were randomly selected for the matched pair, and the source examinee's response vector was overwritten on the copier examinee's response vector for the selected items. This process was repeated until 10 pairs of examinees were obtained to represent each of 12 conditions (copier ability \times source ability \times amount of copying), yielding a total of 120 answer copying pairs per replication.
2. Generating honest pairs: After excluding the 240 examinees used to generate answer copying pairs in Step 1, one remaining examinee with low/high ability was randomly selected and then matched with another randomly selected low/high ability examinee from a different test center in the remaining sample. No copying was simulated for these pairs. This process was repeated until 60 pairs of examinees were obtained to represent each of four conditions (copier ability \times source ability), yielding a total of 240 honest pairs per replication.
3. Computing indices: For the nominal response datasets, item parameters and ability parameters were estimated using the NRM from the datasets contaminated with answer copying. It would not be appropriate to use the NRM item parameters reported in Table A.1 because those item parameters would not represent the dataset obtained after answer copying was simulated for 120 pairs of examinees at each replication. Therefore, the NRM item and person parameters were reestimated after each replication to be used for computing the IRT-based indices. Response similarity indices (Angoff's B , Saupe's t , ω , GBT, K , K_1 , K_2 , S_1 , S_2) were computed for each answer copying and honest pair obtained in Step 1 and 2.² For each dichotomous dataset, item parameters and ability parameters were reestimated using the 2-PL IRT model. Person-fit indices (l_z^* , H^T , and D_0) and response similarity indices (ω , GBT, K , K_1 , K_2 , S_1 , S_2) were computed for each answer copying and honest pair obtained in Step 1 and 2.³

The area under the ROC curve (AUC) was used as an outcome measure to assess the performance of response similarity and person-fit indices under manipulated conditions. In this context, ROC curves plot false positives (the proportion of falsely detected honest pairs) and true positives (the proportion of truly detected answer copying pairs) in a two-dimensional space, and the area under the curve is used as a measure of the classification accuracy for how well an index differentiates the answer copying

pairs from the honest pairs of examinees. The area under the curve for a perfect classification would be equal to one indicating that the index under investigation performs perfectly in classifying answer copying and honest pairs. On the other hand, the area under the curve would be .5 for a chance success in classification, indicating that the index under investigation does not perform beyond chance in classifying answer copying and honest pairs. An area of .5 to .7, .7 to .8, .8 to .9, and .9 to 1 were considered as indicators of poor, fair, good, and excellent classification performance, respectively, in the current study.

Five hundred replications were run. The number of replications was chosen such that the standard error of AUC is always below 0.005, using the formula provided by Hanley and McNeil (1982). As a result of 500 replications, the total number of answer copying pairs was equal to 5,000 (500×10), and the total number of honest pairs was 30,000 (500×60), used for the ROC analysis to estimate the AUC under every condition studied. With a number of 5,000 true positives and 30,000 false positives, the maximum possible standard error for an AUC estimate was equal to 0.0045.

In addition to AUC, the type I error rate of all indices were evaluated at the alpha level of .001, .005, .01, .05, and .10, and the empirical power of all indices were reported for the alpha level of .01. All the steps described in this section were executed by a written R (R Core Team, 2014) routine. Within this R routine, an R package, CopyDetect (Zopluoglu, 2013), was used to compute the response similarity indices. A sample of this routine is given in Appendix B, showing the process from initial data manipulation and exploration to computing these indices for a sample pair of examinees.

Results

The type I error rates of each index were computed by finding the proportion of falsely identified pairs out of the 120,000 honest pairs generated across all conditions. These data are shown in Figure 2.1 for dichotomous and nominal response outcomes. For dichotomous response outcomes, the empirical type I error rates were below the nominal level for all response similarity indices. Among response similarity indices, the empirical type I error rates were closest to the nominal levels for the ω index, while the GBT index and K and its variants seemed to be more conservative. The empirical type I error rates were higher than the nominal levels for the l_z^* . Although the optimal thresholds for the H^T and D_0 were provided in previous research, they are likely to be different for this specific dataset. The threshold values providing a certain type I error rate could be obtained from the first and fifth percentiles of the H^T and 95th and 99th percentiles of the D_0 values computed for all honest pairs simulated across all conditions. For this specific dataset, the values of .045 and 0.084 for the H^T statistic and the values of .223 and .245 for the D_0 statistic were found to be the thresholds corresponding to the empirical type I error rates of 1% and 5%, respectively. For nominal response outcomes, Angoff's B statistic and the procedure proposed by Saupe were found to have significantly inflated type I error rates. The empirical type I error rates were below the nominal level for other response similarity indices, and the ω index provided a better control of type I error rate while the GBT index and K and its variants were similarly found to be more conservative.

A typical practice in answer copying studies is to report the empirical power by computing the proportion of truly identified answer copying pairs within each condition. Therefore, for the sake of completeness, the proportions of truly identified answer copying pairs at the alpha level of .01 are reported in Table 2.3. The readers should

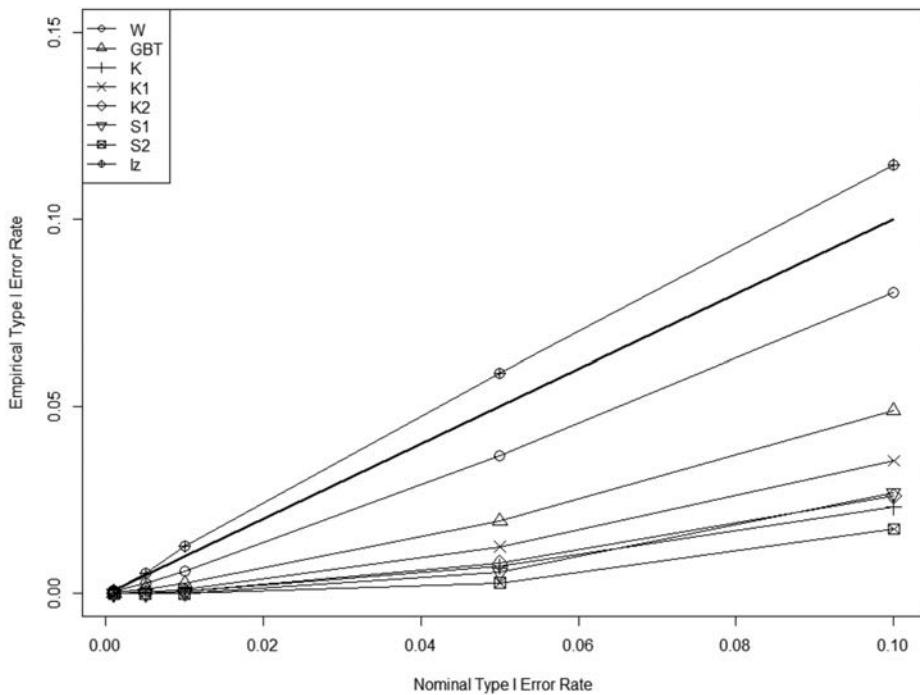


Figure 2.1a Empirical type I error rates of indices for dichotomous and nominal response outcomes. (a) Dichotomous Response Outcome

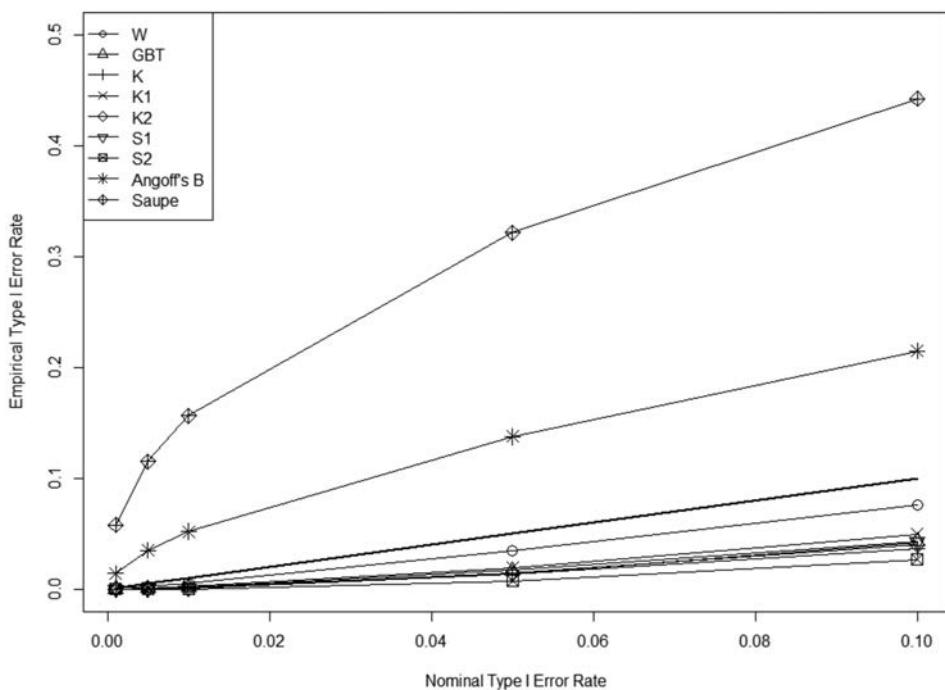


Figure 2.1b Empirical type I error rates of indices for dichotomous and nominal response outcomes. (b) Nominal Response Outcome

Table 2.3 Empirical Power for Simulated Conditions ($\alpha = .01$)

Source Ability	Copier Ability	Amount of Copying	ω	GBT	K	K_1	K_2	S_1	S_2	I_z^*	H^T	D_0
Dichotomous Dataset												
Low	Low	20%	0.16	0.11	0.02	0.03	0.02	0.01	0.01	0.01	0.01	0.10
High	Low	20%	0.05	0.03	0.02	0.03	0.02	0.00	0.00	0.02	0.02	0.09
Low	High	20%	0.22	0.10	0.02	0.03	0.02	0.01	0.00	0.02	0.01	0.00
High	High	20%	0.10	0.04	0.02	0.03	0.03	0.01	0.00	0.01	0.01	0.00
Low	Low	40%	0.72	0.61	0.27	0.39	0.29	0.18	0.11	0.01	0.01	0.11
High	Low	40%	0.34	0.25	0.19	0.28	0.21	0.08	0.04	0.02	0.01	0.05
Low	High	40%	0.81	0.63	0.30	0.37	0.31	0.20	0.12	0.02	0.02	0.01
High	High	40%	0.49	0.32	0.23	0.28	0.27	0.09	0.04	0.01	0.01	0.00
Low	Low	60%	0.98	0.96	0.86	0.92	0.89	0.77	0.67	0.01	0.01	0.12
High	Low	60%	0.80	0.73	0.68	0.75	0.71	0.45	0.29	0.02	0.02	0.02
Low	High	60%	0.99	0.97	0.89	0.93	0.90	0.81	0.73	0.02	0.02	0.04
High	High	60%	0.88	0.79	0.72	0.76	0.76	0.45	0.32	0.01	0.01	0.00
Nominal Dataset												
			ω	GBT	K	K_1	K_2	S_1	S_2	B	<i>Saupe</i>	
Low	Low	20%	0.26	0.11	0.16	0.22	0.17	0.17	0.11	0.16	0.33	
High	Low	20%	0.05	0.03	0.07	0.11	0.08	0.05	0.02	0.08	0.20	
Low	High	20%	0.44	0.13	0.21	0.24	0.21	0.17	0.10	0.08	0.20	
High	High	20%	0.19	0.11	0.08	0.11	0.10	0.05	0.02	0.05	0.15	
Low	Low	40%	0.87	0.72	0.79	0.85	0.81	0.79	0.71	0.30	0.50	
High	Low	40%	0.45	0.37	0.44	0.52	0.48	0.37	0.25	0.13	0.28	
Low	High	40%	0.94	0.80	0.85	0.87	0.85	0.82	0.75	0.13	0.28	
High	High	40%	0.71	0.57	0.50	0.54	0.54	0.38	0.27	0.07	0.18	
Low	Low	60%	0.99	0.98	0.98	0.99	0.98	0.98	0.97	0.47	0.69	
High	Low	60%	0.91	0.89	0.86	0.89	0.88	0.80	0.70	0.21	0.39	
Low	High	60%	0.99	0.98	0.99	0.99	0.99	0.99	0.98	0.19	0.37	
High	High	60%	0.95	0.95	0.88	0.89	0.90	0.81	0.72	0.08	0.22	

Note: The threshold values used for the H^T and D_0 statistics are .045 and .223, respectively, and were determined from the 120,000 honest pairs of response vectors.

interpret them with caution when making comparisons across methods because these methods are not operating at the same empirical alpha level. A more natural measure would be AUC, because ROC analysis allows empirical type I error and power rates to be integrated into one chart, and AUC provides a common scale for all methods under investigation. Therefore, AUC was the main outcome of interest for the current study.

The AUC estimates for the simulated conditions for each index are presented in Table 2.4. An initial exploration when the outcome is dichotomous revealed that the person-fit indices did not perform as well as the response similarity indices in terms of classification accuracy as measured by AUC. Among person-fit indices, D_0 performed relatively better and showed acceptable performance for the conditions in which a low-ability examinee copied 40% or 60% of the items from a high-ability examinee.

Table 2.4 Area Under the ROC Curve for Simulated Conditions

Source Ability	Copier Ability	Amount of Copying	ω	GBT	K	K_1	K_2	S_1	S_2	I_z^*	H^T	D_θ
Dichotomous Dataset												
Low	Low	20%	0.85	0.84	0.85	0.85	0.85	0.85	0.85	0.52	0.50	0.52
High	Low	20%	0.79	0.78	0.79	0.79	0.79	0.78	0.78	0.53	0.54	0.57
Low	High	20%	0.88	0.88	0.88	0.89	0.88	0.89	0.89	0.52	0.57	0.68
High	High	20%	0.82	0.81	0.81	0.81	0.81	0.80	0.80	0.50	0.50	0.51
Low	Low	40%	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.51	0.51	0.53
High	Low	40%	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.56	0.54	0.67
Low	High	40%	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.52	0.59	0.80
High	High	40%	0.96	0.96	0.96	0.95	0.96	0.95	0.95	0.50	0.51	0.50
Low	Low	60%	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.51	0.51	0.54
High	Low	60%	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.57	0.55	0.77
Low	High	60%	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.51	0.59	0.88
High	High	60%	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.50	0.51	0.50
Nominal Dataset												
Source	Copier	Amount	ω	GBT	K	K_1	K_2	S_1	S_2	B	<i>Saupe</i>	
Low	Low	20%	0.91	0.90	0.91	0.91	0.91	0.91	0.91	0.63	0.61	
High	Low	20%	0.86	0.85	0.83	0.83	0.83	0.83	0.83	0.58	0.57	
Low	High	20%	0.93	0.92	0.94	0.94	0.94	0.94	0.94	0.58	0.57	
High	High	20%	0.86	0.84	0.86	0.86	0.86	0.85	0.86	0.55	0.54	
Low	Low	40%	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.74	0.72	
High	Low	40%	0.98	0.98	0.97	0.96	0.97	0.96	0.96	0.66	0.64	
Low	High	40%	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.66	0.64	
High	High	40%	0.98	0.98	0.98	0.97	0.98	0.97	0.97	0.58	0.57	
Low	Low	60%	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.82	
High	Low	60%	1.00	1.00	0.99	0.99	1.00	0.99	0.99	0.73	0.72	
Low	High	60%	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.72	0.71	
High	High	60%	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.63	0.61	

All response similarity indices showed very similar performance for dichotomous response outcomes. When the outcome is nominal, Angoff's B and Saupe's t performed similar to each other and showed acceptable performance for the conditions in which a low ability examinee copied 40% or 60% of the items from a low-ability examinee. The response similarity indices all performed relatively higher than the Angoff's B and Saupe's t methods and performed similarly to each other. To understand the important effects, ANOVA analysis with four between-subjects factors (source ability, copier ability, amount of copying, type of response outcome) and one within-subjects factor (response similarity indices) was run. The main effects for the amount of copying, source ability level, type of outcome, and type of index were all statistically significant ($p < .01$), and the effect sizes, as measured by η^2 , were .805, .072, .031, .007, and .001, respectively. The main effect for the copier ability was not found to be significant, $p = .226$. The only significant two-way interaction was between the type of index

and source ability, but the effect was negligible, $\eta^2 < .001$. The most important factors appeared to be the amount of copying, ability of source examinee, and type of response outcome because the main effects for these three factors explained about 91% of the variance. The average AUC estimate was equal to .996 when the amount of copying was 60% and significantly higher than the average AUC estimate when the amount of copying was 20%, $\Delta\text{AUC} = 0.137, p < .001$. The average AUC estimate was equal to .975 when the amount of copying was 40% and significantly higher than the average AUC estimate when the amount of copying was 20%, $\Delta\text{AUC} = 0.117, p < .001$. There was no significant difference between the average AUC estimates for 40% and 60% copying conditions. The average AUC estimate was equal to .961 when the source was a low-ability examinee and significantly higher than the average AUC estimate when the source was a high-ability examinee, $\Delta\text{AUC} = 0.036, p < .01$. The average AUC estimate was equal to .955 when the nominal response outcomes were used and significantly higher than the average AUC estimate when the response outcome was dichotomous, $\Delta\text{AUC} = 0.024, p < .05$.

ANALYSIS OF FLAGGED INDIVIDUALS AND TEST CENTERS

There were a number of test takers and centers initially flagged by the test provider. This provides an opportunity to test the effectiveness of analytical methods for real cases. In this section, the results from the overall test center integrity analysis and the results from the analysis of flagged individuals are presented. For the ease of reporting, the results only from the ω index were reported. The ω index was chosen because it provided the empirical type I error rates closest to the nominal levels. For the following analysis in this section, the NRM item parameters and person parameters for the 75 common items were obtained from all 3,280 test takers, including the initially flagged 94 test takers, and subsequently used for further analysis, summarized below.

Test Center Integrity

One potential use of the response similarity indices may be assessing test center integrity. There are $n^*(n-1)/2$ number of pairs in a test center. If a response similarity index is applied to all possible pairs in a test center, the proportion of identified pairs should not be exceeding the nominal type I error rate, given that none of the response similarity indices inflate type I error rate based on the literature and the simulation study demonstrated above. For instance, there are 190 possible pairs if a test center has 20 test takers. Given that the ω index is directional and two ω indices can be computed for a pair of test takers, there are 380 statistical tests available for all possible pairs in this test center. If we use a nominal alpha level of .01 to identify a pair, we would expect to find about four significant results just by chance. Suppose you find 20 pairs identified instead of four—this would be an indicator of unusual degree of similarity among test takers in the test center and may warrant additional scrutiny.

There were 326 centers in the current dataset, and the number of test takers in these test centers ranged from 1 to 49, with an average of 10.06. Using the nominal response data, the ω index was computed for each possible pair twice in each test center. Then, the proportion of pairs with significant ω index at the alpha level of .01 and .05 was computed for each test center. Figure 2.2 presents a basic scatter plot of these proportions at the alpha level of .01 and .05. The dashed lines at $x = .02$ and $y = .10$ are the boundaries for flagging test centers, and the choice of these boundaries is somewhat

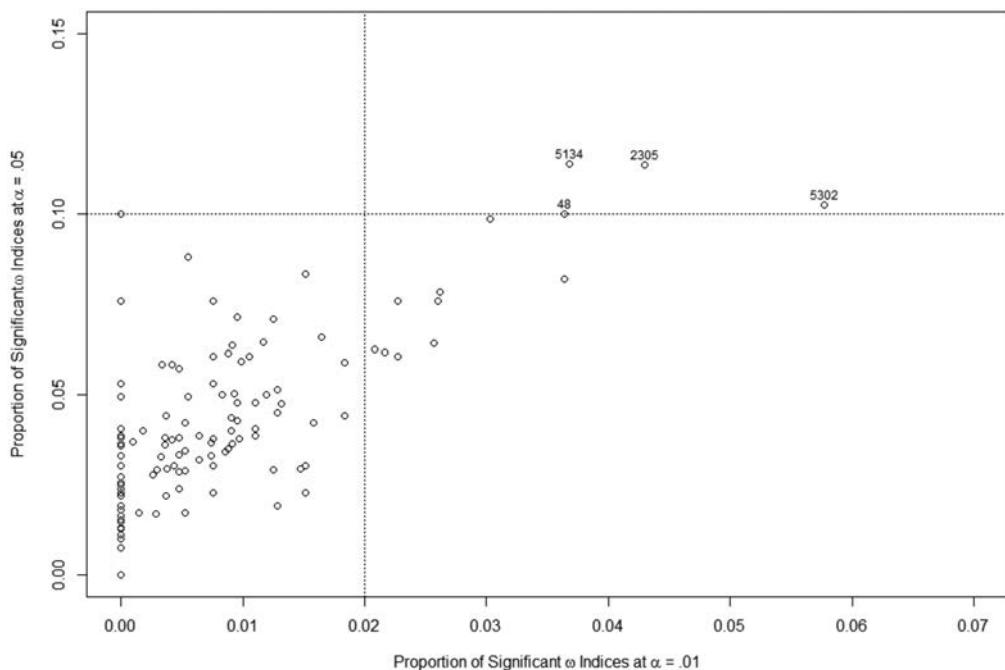


Figure 2.2 Flagged test centers based on the proportion of identified pairs

arbitrary. In the current demonstration, these values were chosen because the author thinks that a proportion twice the nominal type I error rate is very unlikely to be observed. The test centers at the upper right panel in this plot were highly suspicious because the proportion of significant ω indices computed for all pairs in these test centers are beyond the chance level. For instance, there were 49 test takers in the test center 2305, yielding a total number of 1,176 possible pairs and 2,352 ω indices computed. Although the expected number of significant ω indices was about 24 and 118 at the alpha levels of .01 and .05, respectively, the corresponding number of significant ω indices was 101 and 267, indicating an unusual number of pairs identified beyond chance. Not surprisingly, 12 out of 94 test takers flagged initially by the test provider were at test center 2305, confirming this finding.

Analysis of Flagged Individuals

Supplemental analysis was also run for the 94 test takers initially flagged by the test provider. A flagged individual's response vector was compared to all other test takers in the test center. A flagged individual was treated as a copier examinee, and all other test takers were treated as potential source examinees. Based on the findings from the simulation study, the ω index was computed between the flagged individual and other test takers in the same test center and evaluated after using a Bonferroni adjustment depending on the number comparisons for the flagged individual, as suggested by Wollack, Cohen, and Serlin (2001). That is, if the number of comparisons is 10 for a test taker in a test center, each comparison was evaluated at the alpha levels of 0.001 and 0.005, corresponding to the family-wise alpha levels of 0.01 and 0.05 for all 10 comparisons, respectively. The results from the analysis of flagged individuals were reported

Table 2.5 Analysis of Flagged Individuals

ID	Center	N	Number of Flagged Comparisons ($p < .01$)	Number of Flagged Comparisons ($p < .05$)
e100453	556	19	1	1
e101620	1900	7	0	1
e100452	2305	49	0	1
e100505	2305	49	2	3
e100624	2305	49	4	6
e200336	2305	49	0	2
e200417	2305	49	1	1
e200448	2305	49	1	1
e200503	2305	49	0	1
e100494	5302	13	0	2
e100524	5856	20	0	1

Note: Number of total comparisons for each test taker is equal to the number of other test takers in the same test center ($N-1$). The alpha level for each comparison were adjusted using a Bonferroni correction based on the number of comparisons to maintain the family-wise alpha levels of .01 and .05.

in Table 2.5. For instance, the response vector of the test taker with ID “e100624” was compared to all other 48 test takers in test center 2305. Out of 48 comparisons, four were found to be significant at the alpha level of .00021 (.01/48), and six were found to be significant at the alpha level of 0.00104 (.05/48). Table 2.5 includes the information about test takers with at least one significant comparison at the family-wise alpha level of .05. As seen in Table 2.5, the ω index yielded at least one significant result for 11 test takers at the family-wise alpha level of .05 and five test takers at the family-wise alpha level of .01 out of 94 initially flagged test takers. Yet more, the ω index supported the fact that there was some sort of systematic test fraud in test center 2305 because the ω index identified unusual response similarity for seven test takers with a number of other individuals in the same test center.

CONCLUSIONS

The integrity of test scores will continue to be a major concern as long as test-based accountability remains to be a part of the educational system. Although there are administrative procedures and methods to proactively prevent testing irregularities, the use of comprehensive integrity analyses after every test administration is highly recommended for investigating different types of potential test fraud in academic and certification/licensure testing (NCME, 2012; U.S. Department of Education, IES, & NCES, 2013). Not surprisingly, legislators in some states were pushing bills that required state departments to regularly look for potential fraud in standardized tests after every administration (Crouch, 2012), while some other states created test security units within their departments of education (Hildebrand, 2012). This chapter attempted to provide some insights to practitioners on the existing quantitative methods that have been developed to detect and combat answer copying/sharing through

the use of both a simulation study and a demonstration of the use of these methods in a real test administration with flagged test takers and centers.

The simulation study suggested that person-fit indices included in this study do not perform sufficiently well in identifying examinees engaged in copying/sharing answers. Two previous studies (de la Torre & Deng, 2008; Karabatsos, 2003) found relatively better performances for the person-fit indices included in this study. Relatively lower performance of these methods in the current study can be attributable to the differences in study design. For instance, Karabatsos (2003) simulated answer copying by generating item responses for a low-ability examinee with an ability range of [-2, -.5] and imputing correct responses for the 18% of the most difficult items. It is not surprising that person-fit indices perform relatively better when a low ability examinee gives correct responses to a certain percentage of most difficult items. However, answer copying/sharing may contain a broader range of response changing patterns in a real test administration, and sharing/copying answers from another student does not necessarily yield an aberrant response vector. The person fit-indices may be useful for identifying other types of aberrance, such as anxiety, lack of motivation, carelessness, or lack of proficiency in language. It is also possible that person-fit indices may be useful for detecting other types of test compromise, such as item preknowledge. However, they do not appear useful for the detection of answer copying/sharing. One limitation to note for the simulation study is the ratio of response vectors contaminated with answer copying. The percentage of contaminated response vectors was about 3.8% in the current design. Item parameter estimates would be affected by the amount of contamination and may have an indirect effect on power and type I error rates of these indices through more accurately or inaccurately estimated item parameters for a lower or higher percentage of contaminated response vectors.

A number of response similarity indices provided sufficient performance in detecting unusual agreement between two response vectors. In terms of controlling the nominal type I error rate, the ω index was the most successful method among response similarity indices by providing the closest empirical type I error rates to the nominal levels. It should be noted that the ω index was also most powerful in detecting true copying pairs at a specified alpha level (e.g., .01). However, it should be noted that the lack of power for other indices are due to their more conservative nature with very small empirical type I error rates than nominal levels. While the most important factor influencing the performance of response similarity indices was found to be the amount of copying, the ability of the source examinee and the type of response outcome also had some effect on performance. The current chapter also demonstrated a potential use of response similarity indices for monitoring test center integrity. Given the information that these indices do not inflate the type I error rate, test centers can be flagged for additional scrutiny due to an unusual proportion of identified pairs (e.g., twice as large as nominal type I error rate). The demonstration using the ω index successfully flagged one test center which contained about 13% of initially flagged test takers.

Before closing, it should be noted that all methods included in the current chapter rely on a traditional frequentist approach. In other words, these methods compute the likelihood of a response pattern or agreement between two response vectors given the assumption that a test taker or a pair of test takers are not engaged in sharing/copying responses. Alternatively, van der Linden and Lewis (2015) presented a Bayesian approach to compute the posterior odds of actual copying, which is a relatively untouched area of research. Also, the methods included in the current chapter, at least as they relate to answer copying, are mostly designed for large-scale paper-and-pencil

assessments. Although these methods or underlying ideas can still find potential applications for computer-based testing where test takers receive a same set of items, their utility may be decreased as the use of computerized adaptive testing continues to increase.

NOTES

1. See Buss and Novick (1980) and Dwyer and Hecht (1994) for detailed legal discussions of utilizing evidence for response similarity analysis.
2. Person-fit indices are defined for dichotomous response vectors, so they are excluded for the analysis of nominal data.
3. Angoff's *B* and Saupe's *t* are appropriate to be used with nominal response data, so they are excluded for the analysis of dichotomous data.

REFERENCES

- Angoff, W. H. (1972). *The development of statistical indices for detecting cheaters*. Berkeley, CA: Educational Testing Service.
- Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, 69(345), 44–49.
- Anikeeff, A. (1954). Index of collaboration for test administrators. *Journal of Applied Psychology*, 38(3), 174–177.
- Association of Test Publishers, and National College Testing Association. (2015). *Proctoring best practices*. CreateSpace Independent Publishing Platform.
- Bay, L. (1995, April). *Detection of cheating on multiple-choice examinations*. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.
- Bellezza, F. S., & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error similarity analysis. *Teaching of Psychology*, 16(3), 151–155.
- Belov, D. I. (2011). Detection of answer copying based on the structure of a high-stakes test. *Applied Psychological Measurement*, 35(7), 495–517.
- Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via kullback-leibler divergence and K-index. *Applied Psychological Measurement*, 34(6), 379–392.
- Bird, C. (1927). The detection of cheating in objective examinations. *School and Society*, 25(635), 261–262.
- Bird, C. (1929). An improved method of detecting cheating in objective examinations. *The Journal of Educational Research*, 19(5), 341–348.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1), 29–51.
- Bopp, M., Gleason, P., & Misicka, S. (2001). *Reducing incidents of cheating in adolescence* (master's thesis). Saint Xavier University, Chicago, IL.
- Brimble, M., & Clarke, P. S. (2005). Perceptions of the prevalence and seriousness of academic dishonesty in Australian universities. *Australian Educational Researcher*, 32(3), 19–44.
- Buss, W. G., & Novick, M. R. (1980). The detection of cheating on standardized tests: Statistical and legal analysis. *Journal of Law and Education*, 6(1), 1–64.
- Cai, L., du Toit, S. H. C., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago, IL: Scientific Software International.
- Cody, R. P. (1985). Statistical analysis of examinations to detect cheating. *Journal of Medical Education*, 60(2), 136–137.
- Crouch, E. (2012, April 4). School test fraud is target of Missouri bill. *St. Louis Today*. Retrieved from www.stltoday.com/news/local/education/school-test-fraud-is-target-of-missouri-bill/article_ab95386b-48c2-50ef-b5c6-e110a445019b.html
- De La Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, 45(2), 159–177.
- Dickenson, H. (1945). Identical errors and deception. *Journal of Educational Research*, 38(7), 534–542.
- Dimitrov, D. M., & Smith, R. M. (2006). Adjusted Rasch person-fit statistics. *Journal of Applied Measurement*, 7(2), 170.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1), 67–86.

- Dwyer, D. J., & Hecht, J. B. (1994). *Cheating detection: Statistical, legal, and policy implications*. Normal, Illinois: Illinois State University.
- Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, 2(4), 235–256.
- Gregory, R. J. (2004). *Psychological testing: History, principles, and applications*. Boston, MA: Allyn & Bacon.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36.
- Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying*. (ACT Research Report No. 87-15). Iowa City, IA: American College Testing.
- Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18, 133–146.
- Hildebrand, J. (2012, March 15). State creates unit to combat test fraud. *Newsday*. Retrieved from www.newsday.com/long-island/education/state-creates-unit-to-combat-test-fraud-1.3605217
- Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-Index: Statistical theory and empirical support* (ETS Research Report No. 96-97). Princeton, NJ: Educational Testing Service.
- Hughes, J. M. C., & McCabe, D. L. (2006). Academic misconduct within higher education in Canada. *Canadian Journal of Higher Education*, 36(2), 1–21.
- Jensen, A. L., Arnett, J. J., Feldman, S. S., & Cauffman, E. (2002). It's wrong, but everybody does it: Academic dishonesty among high school and college students. *Contemporary Educational Psychology*, 27(2), 209–228.
- Josephson Institute of Ethics. (2006). *The ethics of American youth*. Retrieved from <http://charactercounts.org/pdf/reportcard/2006/reportcard-all.pdf>
- Josephson Institute of Ethics. (2008). *The ethics of American youth*. Retrieved from http://charactercounts.org/pdf/reportcard/2008/Q_all.pdf
- Josephson Institute of Ethics. (2010). *The ethics of American youth*. Retrieved from http://charactercounts.org/pdf/reportcard/2010/ReportCard2010_data-tables.pdf
- Josephson Institute of Ethics. (2012). *The ethics of American youth*. Retrieved from <http://charactercounts.org/pdf/reportcard/2012/ReportCard-2012-DataTables.pdf>
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16(4), 277–298.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational and Behavioral Statistics*, 4(4), 269–290.
- Lin, C. H. S., & Wen, L. Y. M. (2007). Academic dishonesty in higher education: A nationwide study in Taiwan. *Higher Education*, 54(1), 85–97.
- McCabe, D. (2001). Cheating: Why students do it and how we can help them stop. *American Educator*, 25(4), 38–43.
- Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, 18(2), 111–120.
- National Council on Measurement in Education. (2012). *Testing and data integrity in the administration of statewide student assessment programs*. Retrieved from <http://ncme.org/default/assets/File/Committee%20Docs/Test%20Score%20Integrity/Test%20Integrity-NCME%20Endorsed%20%282012%20FINAL%29.pdf>
- Olson, J., & Fremer, J. (2013). *TILSA test security guidebook: Preventing, detecting, and investigating test security irregularities*. Washington, DC: Council of Chief State School Officers.
- R Core Team. (2014). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.
- Rakovski, C. C., & Levy, E. S. (2007). Academic dishonesty: Perceptions of business students. *College Student Journal*, 41(2), 466–481.
- Saretzky, G. D. (1984). *The treatment of scores of questionable validity: The origins and development of the ETS Board of Review* (ETS Occasional Paper). Princeton, NJ: Educational Testing Service. Retrieved from <http://files.eric.ed.gov/fulltext/ED254538.pdf>
- Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo: Meiji Tokyo.
- Saupe, J. (1960). An empirical model for the corroboration of suspected cheating on multiple-choice tests. *Educational and Psychological Measurement*, 20(3), 475–489.
- Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden*, 7, 131–145.
- Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's non-parametric IRT model. *Applied Psychological Measurement*, 16(2), 149–157.
- Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the K-index for detecting answer copying. *Journal of Educational Measurement*, 39(2), 115–132.
- Sotaridona, L. S., & Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement*, 40(1), 53–69.

- Thissen, D., Chen, W. H., & Bock, R. D. (2003). Multilog (Version 7) [Computer software]. Lincolnwood, IL: Scientific Software International.
- Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 83–108). New York, NY: Academic Press.
- U.S. Department of Education, IES, & NCES. (2013). *Testing integrity symposium: Issues and recommendations for best practice*. Retrieved from <http://nces.ed.gov/pubs2013/2013454.pdf>
- Vandehey, M. A., Diekhoff, G. M., & LaBeff, E. E. (2007). College cheating: A twenty-year follow up and the addition of an honor code. *Journal of College Student Development*, 48(4), 468–480.
- van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties* [Comparability of individual test performance]. Lisse: Swets & Zeitlinger.
- van der Linden, W. J., & Lewis, C. (2015). Bayesian checks on cheating on tests. *Psychometrika*, 80(3), 689–706.
- van der Linden, W. J., & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, 31(3), 283–304.
- Wesolowsky, G. O. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics*, 27(7), 909–921.
- Whitley, B. E. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education*, 39(3), 235–274.
- Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, 21(4), 307–320.
- Wollack, J. A. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement*, 40(3), 189–205.
- Wollack, J. A. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education*, 19(4), 265–288.
- Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement*, 22(2), 144–152.
- Wollack, J. A., Cohen, A. S., & Serlin, R. C. (2001). Defining error rates and power for detecting answer copying. *Applied Psychological Measurement*, 25(4), 385–404.
- Zopluoglu, C. (2013). CopyDetect: an R package for computing statistical indices to detect answer copying on multiple-choice examinations. *Applied Psychological Measurement*, 37(1), 93–95.
- Zopluoglu, C., & Davenport, E. C., Jr. (2012). The empirical power and type I error rates of the GBT and ω indices in detecting answer copying on multiple-choice tests. *Educational and Psychological Measurement*, 72(6), 975–1000.

3

DETECTING POTENTIAL COLLUSION AMONG INDIVIDUAL EXAMINEES USING SIMILARITY ANALYSIS

Dennis D. Maynes

INTRODUCTION

This chapter demonstrates how potential collusion among individual examinees using similarity statistics may be detected. Most of the research in this area has been conducted using answer-copying statistics (Frary, Tideman, & Watts, 1977; Wollack, 1997). These statistics have generally been used to determine whether a person, known as the copier, potentially copied answers from another person, known as the source. However, this approach is limited in practice because it does not account for two or more test takers working together to breach the security of the exam by communicating and sharing test content and/or answers to questions during the exam. The source-copier approach also does not provide a mechanism to detect collusion that resulted from communication external to the testing session. In exchange for setting aside the assignment of a potential source and copier, similarity statistics are able to detect potential collusion in these and other situations. Although answer copying and similarity statistics are related, the focus in this chapter will be entirely on answer similarity statistics. A more complete discussion of answer copying statistics can be found in Zopluglu (this volume).

With the pervasive use of technology for testing and for cheating, cheating by copying answers is no longer the primary threat that it once was. Indeed, electronic devices have been used by test takers to communicate and share answers in real time across great distances. Hence, similarity statistics provide a means of detecting general forms of collusion that are not as easily detected by answer-copying statistics. In recent years, researchers have been exploring these more general forms of potential test fraud (Zhang, Searcy, & Horn, 2011; Belov, 2013).

PROPERTIES OF SIMILARITY STATISTICS

In published research for answer-copying and similarity statistics, most researchers have restricted their work to pairs of students where potential answer-copying was

present. The more general test security threat of collusion has not received much attention by academic researchers. Collusion describes a testing scenario where two or more examinees are working together for the personal gain of at least one of those examinees. Answer copying is a special case of collusion for which the source examinee may not realize his or her own involvement. However, collusion also extends to cover instances where examinees are deliberately sharing their answers; communicating during the exam, verbally or nonverbally; working together before the exam to share content (i.e., acquire preknowledge); using surrogate or proxy test takers; or receiving disclosed answers to test questions by instructors, teachers, and trainers. This chapter focuses on the role of similarity statistics in helping to detect test collusion. Wollack and Maynes (this volume) describe an application of the approach discussed here for purposes of detecting and extracting groups of examinees among whom collusion was potentially present.

Answer copying statistics require specifying both a copier and a source, and because the statistic is conditional on the responses provided by the source examinee, the value of the statistic is different depending on which examinee is treated as the copier and which as the source. In contrast, similarity statistics have the property of symmetry; that is, they produce the same values regardless of whether one examinee is a source and another is an information beneficiary. Symmetry is a desirable property because a copier does not need to be hypothesized and because it allows for clustering techniques to be used to extract a group structure.

Another desirable property of similarity statistics is that their probability distributions are based on the assumption of independence, hence allowing for a specific statistical hypothesis of independent test taking to be tested. The term *nonindependent test taking*, instead of *cheating* or *collusion*, is used in this chapter describing the scenario when the hypothesis of independent test taking is rejected, even though similarity statistics are designed to detect potential cheating and/or collusion.

One point of debate among similarity (or answer copying) researchers relates to which items should be considered (see Wollack & Maynes, 2011). The heart of this debate is whether or not matches on correct items should be considered as providing evidence of misconduct. Clearly, each identical incorrect answer provides some evidence of wrongdoing, and many indexes focus exclusively on matches of this variety (Holland, 1996; Bellezza & Bellezza, 1989). However, others have argued that valuable information can also be extracted from correct answers, even if not as much, and so have designed indexes that consider all items (Wesolowsky, 2000; Wollack, 1997; van der Linden & Sotaridona, 2006).

It is the opinion of this author that using the entire set of test responses is preferred because the outcome of a response may be viewed as a random event (under assumptions of local independence within IRT models). When the similarity statistic is based on identical incorrect responses only, a critical piece of evidence is ignored. The purpose of cheating on tests is to get a high score. This is only possible when identical correct responses are present. For example, suppose two or more test takers have gained access to a disclosed answer key containing most, but not all, of the answers. Statistics that focus exclusively on identical incorrect responses will likely not detect this situation (assuming that the incorrect responses were from items with nondisclosed answers). Statistically, the probability distribution of identical incorrect responses depends upon the number of identical correct responses (see Table 3.2 in this chapter for this exposition). The number of identical correct responses may be extreme, but when it is used to condition the probability distribution (i.e., when it is treated as an

variable not providing information about potential test fraud), the extreme improbability of observed similarities will not be accurately portrayed.

MAKING STATISTICAL INFERENCES WITH SIMILARITY STATISTICS

There is some debate among practitioners concerning the specific statistical inferences that should be made with respect to collusion on tests. Statistical inferences can and should be made about test score validity (i.e., can you trust the score?). Statistical inferences can also be made about test-taker behavior (i.e., did the person cheat?). There seems to be some confusion among researchers about which inferences can properly be made and how they should be made. Regardless of disagreements among researchers, proper statistical inferences require modeling the statistical distribution of the similarity statistic under the assumption that tests were taken normally (i.e., no cheating or testing irregularity occurred). Thus, the distribution of the similarity statistic needs to be modeled before statistical inferences may be made.

Discussion of Some Similarity Statistics

Because this chapter is concerned with similarity statistics, common answer-copying statistics such as ω (Wollack, 1997), g_2 (Frary et al., 1977), and the K index (Holland, 1996) are not discussed. Instead, a few salient similarity statistics are now discussed in chronological order.

Hanson, Harris & Brennan (1987) described two similarity statistics, PAIR1 and PAIR2. These are bivariate statistics, with PAIR1 equal to the number of identical incorrect responses *and* length of the longest string of identical responses, and PAIR2 equal to the number of identical incorrect responses in the longest string of identical responses *and* the ratio of the number of identical incorrect responses to the sum of the nonmatching responses and the number of identical incorrect responses. A theoretical, model-based method for estimating the distributions for these statistics was not provided by the authors. Instead, they suggested that empirical sampling could be used to assess the extremity of a particular pair of similar tests. As a result, probability estimates of observed similarities for these two statistics are not currently computable. Because tail probabilities are not computable for these statistics, they are not suitable to use for detection because appropriate Type I error control cannot be imposed. However, they have been and will likely continue to be used for purposes of supporting allegations of testing improprieties.

Bellezza & Bellezza (1989) introduced error similarity analysis (ESA). The statistic in this analysis is the number of incorrect matching responses between a pair of test takers. The distribution is based on a binomial distribution where the conditional probability of providing a matching wrong answer was estimated by counting the total number of matching wrong answers and dividing by the total number of shared wrong answers in the data set. The same authors published an update (1995) in which they suggested that the single conditional probability estimate for conditionally matching incorrect answers could be replaced with estimates for each item on the test, if desired.

Wesolowsky (2000) introduced a statistic with the probability calculated by a computer program, S-Check. This statistic counts the number of matching answers for a pair of test takers. The probability of a matching answer is computed using the assumption of independence and the performance level of the two takers, where probabilities

of correct answers are estimated by a smooth function “suggested by l_p distance iso-contours from location theory” (p. 912). The probability that student j will correctly answer question i is

$$p_{ij} = [1 - (1 - r_i)^{a_j}]^{1/a_j},$$

where a_j is a parameter that estimates student ability (similar to the function of θ in item response theory) and r_i is equal to the proportion of students that answered the item correctly (i.e., the item’s p -value). Probabilities of incorrect matching answers are computed using the product rule (i.e., assuming independence), and probabilities of individual incorrect responses conditioned upon an incorrect response (see Wesolowsky’s paper for details). Wesolowsky states that the number of matching answers follows a generalized binomial distribution, but for sake of computational efficiency he approximated the tail probability using the normal distribution.

Van der Linden & Sotaridona (2006) introduced a similarity statistic, GBT, based on the generalized binomial distribution. This statistic counts the total number of matching answers. The probability of a matching answer is computed using the assumption of independence and the performance level of the two takers, where response probabilities are estimated using the nominal response model (Bock, 1972).

Maynes (2014) described a bivariate statistic, M4, which consists of the number of identical correct and the number of identical incorrect answers. Following van der Linden and Sotaridona, Maynes uses the nominal response model to estimate item response probabilities conditioned upon test taker performance. The probability distribution for this statistic is postulated to follow a generalized trinomial distribution, where probabilities of matching responses are estimated using the assumption of statistical independence and item response probabilities from the nominal response model.

Violations of Assumptions

The derivations of the distributions of these statistics assume (1) responses between test takers are stochastically independent; (2) response probabilities depend upon test-taker performance, which can be modeled using a mathematical model; and (3) item responses are “locally” or conditionally independent and only depend upon test-taker performance. As seen in the previous section, several models have been used by researchers, with the nominal response model being referenced and used most recently. The choice of model will depend upon the data and the computational tools that are available for estimating the model. Model suitability is always a question that should be asked. When possible, the author prefers to verify model appropriateness through goodness of fit tests or by empirically sampling from live data that demonstrate the extremity of pairs detected by the similarity statistic (e.g., similar to the sampling procedures recommended by Hanson et al., 1987).

Extreme values of similarity statistics provide evidence of nonindependent test taking (i.e., some common characteristic or occurrence shared by two individuals has influenced their answer selections). Observed similarity between the test responses (or nonindependent test taking) can result from several situations or factors (including a few nonfraudulent behaviors), most of which are described in this section. Even though the similarity statistics are designed to detect nonindependent test taking, additional information (e.g., obtained by examining seating charts or interviews) may be required to determine the nature of the nonindependent test taking that occurred.

While the statistics may not tell us what happened or the behavior that led to nonindependent test taking, they may provide information for making inferences concerning unobserved behaviors and the trustworthiness of test scores. The degree or amount of nonindependence (i.e., deviation from typical or expected similarity), the scope of nonindependence (i.e., the size of the detected clusters), and the pattern of nonindependence among the responses (i.e., the alignment of matching and nonmatching responses) provide important clues as to why the null hypothesis of independent test taking was rejected.

An understanding of the assumptions of the statistical procedures is critical to the practitioner, because a violation of the assumptions may result in observing extreme values of the similarity statistic for two or more test instances. Understanding causes and effects that can be responsible for nonindependent test taking helps the practitioner evaluate possible explanations for an extreme value of the statistic.

It is often the case that similarity statistics demonstrate robustness when the assumptions do not strictly hold. For example, it is seldom true that all test takers in the population are independent. Indeed, many test takers offer “studying together” as a defense against statistical detection of potential test fraud. In offering this “defense,” they seem to overlook the fact that nearly all test takers study together.

Beyond the property of robustness, another important statistical concept is the idea of effect size. When a statistical assumption is violated, one should ask whether the violation would result in a large or small effect. Research suggests that studying together and sharing a common environment will only result in small, not large, effects (Allen, 2014).

Some ways in which assumptions may be violated are now listed. Because robustness of the statistics to violations and the effect sizes of violations are not known, it would be improper to place undue emphasis on them. Also, it is important to remember that these statistics are employed to detect potentially fraudulent test-taking behaviors.

1. The test responses were not independent.
 - a. **Instructional bias:** This can be especially problematic when students are taught to solve problems incorrectly or when the teaching is wrong.
 - b. **Collaborative learning:** There is a definite line between instruction and collusion, but at times that line blurs when students are encouraged to “help” each other.
 - c. **Repeated test taking by the same test taker:** Sometimes test takers are administered the same test again. Such an administration may generate excess similarity if the test taker remembers previous answers to the same question.
 - d. **Test fraud:** The following behaviors, which are generally acknowledged as some form of cheating, violate the assumption of independence:
 - i. Test takers obtain preknowledge of the test questions and/or answers (e.g., through the internet or some other media).
 - ii. Test takers copy from each other during the test.
 - iii. Test takers communicate with each other during the test.
 - iv. Test takers receive assistance or answers to questions from the same helper (e.g., an instructor or teacher).
 - v. Test takers use the same stand-in or proxy to take the test for them.
 - vi. A person responsible for administering the test (e.g., an educator or teacher) changes answers that were provided by the test takers (e.g., by erasing and marking the correct answer).

2. The selected mathematical model does not adequately estimate or approximate the match probabilities.
 - a. **Partial or incomplete tests:** Occasionally test takers omit responses. In general, shared omitted responses do not constitute evidence of potential test fraud. But, a large number of omitted responses can bias the probability estimates, leading to errors in computing the probability of the similarity statistic.
 - b. **Mislabeling of data:** Errors in the data may result in apparent nonindependent test taking. For example, suppose the exam is given with two forms and two test takers are given Form A but their data is labeled as Form B. The extreme unusualness of the response patterns when scored under Form B contributes to extremely improbable alignments, which are perfectly reasonable under Form A.
 - c. **Item exposure or item drift:** When a subset of the items has become exposed to the extent that most test takers are aware of the item content, the items may be easier than previously estimated. This can cause errors in the probability computations.
 - d. **Negative item discriminations:** Items that have negative correlations with the total test score present a unique challenge for models that incorporate test-taker performance. Usually, only very small number of these items are present on exams, but a large number of these items can negatively affect probability computations.
 - e. **Nonmeasurement items:** Sometimes the testing instrument is designed and used to collect information besides the answers to the test questions (e.g., demographic questions or acceptance of a nondisclosure agreement). Unless removed, these items can create unwanted noise in the models and the probability computations.
 - f. **Differential item functioning:** If subgroups exist where the items perform in a distinctly different way than they do for the main population, improbable alignments may occur. If this type of problem is suspected, it may be appropriate to conduct the similarity analysis using subpopulation-specific models.
3. Item responses are not “locally” or conditionally independent.
 - a. **Constant answering strategies and patterned responses:** Two or more test takers use the same answering strategy, which results in them answering the questions in nearly the same way. For example, one of these strategies is “Guess ‘C’ if you don’t know the answer.”
 - b. **Multipart scenario items:** Responses for these items are generally provided sequentially, and responses may depend upon previous responses. This violation can inflate the value of the similarity statistic, yielding false positives.

When anomalous pairs of similar tests are found, it is important to determine the reason. The above discussion has listed potential explanations for extreme values of the similarity statistic. Analysis of live and simulated data indicates that inappropriate increases in test scores and pass rates also may be observed when the similarity is due to nonindependent test taking.

Although care has been taken to list how the assumptions might be violated, it should be remembered that similarity statistics are computed from pairs of data. As a result, in order for a pair to be detected by a similarity statistic, the assumptions must

be violated in the same way for both members of the pair. Because of this, similarity statistics are quite robust to violations of assumptions that do not involve fraudulent manipulation of the test responses.

Discussion of Exploratory and Confirmatory Analyses

An exploratory analysis is conducted when the similarity statistic is computed for all pairs of test instances within the data set, without any preconceived hypotheses about which individuals may have been involved in cheating. Exploratory analyses are synonymous with data mining or data dredging. These are usually done to monitor the test administration for the existence of potential collusion. Because all the pairs have been computed and examined, unless a multiple-comparisons correction is used, a spuriously high number of detections will be observed. The Bonferroni correction or adjustment provides one way to account for conducting multiple related tests (i.e., that an examinee did not work independently on his or her exam).

Typically, the Bonferroni correction is used to establish a critical value for the test of hypothesis. The critical value is that value from the distribution that has a tail probability equal to the desired alpha level divided by the number of elements in the population that were examined (or the number of hypotheses that were statistically evaluated). When exploring all possible pairs, Wesolowsky (2000) recommended using $N \times (N - 1)/2$ as the denominator for the Bonferroni adjustment, because that is how many comparisons were performed.

However, the simple adjustment to the critical value of the similarity statistic suggested by Wesolowsky is not appropriate when multiple data sets of varying sizes (e.g., schools or test sites) are analyzed. In fact, Wesolowsky's suggestion would result in critical values which vary from data set to data set. This is an unsatisfactory solution because all test takers would not be measured against the same standard or threshold.

It is difficult to maintain the same standard for evaluating the pairs in an exploratory analysis, because the recommendation is to perform an adjustment that depends upon N , which varies. If observations are added to a data set, which is then reanalyzed, the standard will change. If, as mentioned above, the number of comparisons is restricted using a proximity measure (e.g., the test takers were tested at the same test site or on the same day), then the standard will vary. Hence, a reasonable approach is needed to acknowledge this situation and at the same time apply a consistent rule for extracting pairs. This can and should be done by ensuring that the multiple comparisons adjustment is applied in a way that takes into account the number of pairwise comparisons and the number of test takers, even when the analysis is restricted to subgroups of varying sizes (e.g., different numbers of test takers were tested at the same test sites or on the same days).

The approach which has confirmed to be satisfactory through simulation is to factor the Bonferroni equation¹ using two terms. The procedure converts the similarity statistic into an “index value” using the relationship, $p = 10^{-I}$, where p is the upper tail probability value of the similarity statistic and I is the index value (Maynes, 2009). In the inverted formulation of the problem, the probability value associated with the Bonferroni critical value becomes $p_{c,s} = p/(N \times (N_s - 1)/2) = \{10^{-I} \times 10^{-\log_{10}((N_s - 1)/2)}\} \times 10^{-\log_{10}(N)}$. Instead of adjusting critical values, this formulation adjusts the similarity statistic (or its probability), thereby taking into account differences in data set sizes and allowing use of the same critical value for every single test taker in the population. The first factor takes into account the number of comparisons for a particular test

taker or student. It determines the index value for the maximum observed similarity statistic (as determined by the smallest tail probability), using the relation $I_{max,s} = I + \log_{10}((N_s-1)/2)$, where N_s is the number of comparisons for a particular student or test taker. At this point, all the index values, $I_{max,s}$, are comparable from student to student. If desired, the second factor can be applied to take into account the number of test takers that were analyzed, $I_s = I_{max,s} + \log_{10}(N)$, where N is the number of students in the analysis.

A confirmatory analysis, on the other hand, is usually conducted on those response strings where independent evidence (e.g., a proctor's testing irregularity report) suggests that test security might have been violated. When performed this way, the Bonferroni adjustment should not be used, unless as suggested by Wollack, Cohen & Serlin (2001) for dealing with specific individuals who are being investigated multiple times, as might be the case if an individual is believed to have copied from multiple neighboring examinees.

There has been some debate among researchers and practitioners concerning the critical value that should be used for a confirmatory analysis. It appears that Angoff (1974) was the first author to report use of a specific critical value. He reported that ETS requires 3.72 standard deviations or greater before taking action, which corresponds with an upper tail probability value of 1 in 10,000 if the distribution of the statistic were normal. Angoff further states that the distribution of the similarity statistic will most likely be skewed and not normal. When asked for a recommended critical value to invalidate a test score, many practitioners will follow Angoff's critical value because (1) it is conservative, (2) similarity may be caused by nonfraudulent factors, and (3) current methods only approximate actual statistical distributions.

However, Angoff's critical value should not be taken as the de facto industry standard. On the contrary, selection of the critical value is a policy decision that is the sole responsibility of each testing organization. The decision should adhere to accepted scientific practice and abide by the organization's goals and responsibilities. Decisions concerning scores must take into account (1) the totality of evidence for individual situations, (2) the effect of Type I errors (i.e., false positives) on test takers balanced against the harm from Type II errors (i.e., failure to take action when nonqualified individuals receiving passing scores), and (3) the organization's ability to implement its policy decisions (e.g., if the number of flagged pairs is excessively burdensome, fewer pairs should be handled). Recommendation of a single critical value is somewhat simplistic because it cannot adequately address these factors.

Limitations of Similarity Statistics and Models

Similarity statistics have the potential to detect and/or expose significant security risks to the exams. Even so, the statistics are subject to some limitations, in addition to situations that may violate the assumptions. These are:

1. A sufficiently large sample size must be used to estimate the match probabilities adequately. The question of adequateness may be answered by goodness-of-fit analyses. The question "How large of a sample is needed?" depends upon model assumptions. For example, a three-parameter logistic model requires a larger sample for estimation than a Rasch model.
2. Similarity statistics cannot determine responsibility and/or directionality. For example, some answer-copying statistics are computed under the hypothesis that

the “source” and the “copier” are known. A similarity statistic is unable to make an attribution of a source or a copier. Additional information is needed to make such an inference.

3. Similarity statistics cannot detect cheating that occurs through some other means than nonindependent test taking. For example, if a surrogate or proxy test taker is employed by only one test taker, the similarity statistic will not be able to detect this.
4. As discussed above, similarity statistics, when used to compare each test with every other test (i.e., data mining), must utilize an approach to adjust the index (or critical) value to reflect multiple comparisons and keep the Type I error rate controlled. However, in doing so, it is important to recognize that the power of the statistical procedure will be lowered.
5. Similarity statistics are sensitive to the number of items in common between two test instances. If two examinees share very few items (e.g., as with a CAT test), the power of similarity statistics to detect collusion between those individuals will generally be low.
6. The distribution of similarity statistics depends upon test-taker performance. Power decreases as the test scores increase. Thus, similarity statistics cannot detect cheating between tests when performance levels are very high (e.g., nearly every question is answered correctly).

DIFFICULTIES IN MODELING THE DISTRIBUTIONS OF SIMILARITY STATISTICS

When Angoff (1974) published his analysis, he dismissed the notion that the distribution of the similarity statistic could be modeled theoretically. He stated, “However, even a brief consideration of this [problem] makes it clear that the complexities in making theoretical estimates of such a distribution are far too great to make it practical” (p. 44). His primary objection was that there was no known way to model the correlations between correct and incorrect responses.

Using assumptions of statistical independence conditioned upon test-taker performance, Item Response Theory (Lord, 1980) is able to model the correlations between the item responses for each pair of test takers. The NRM (Bock, 1972) allows correlations to be modeled for each distinct response. Thus, modern IRT addresses Angoff’s concerns and provides the framework for computing the statistical distribution of the similarity statistics.

It should be emphasized that these distributions cannot be modeled without conditioning upon examinee performance (i.e., θ) and without recognizing that items (and item alternatives) vary in difficulty. For example, the Error Similarity Analysis by Bellezza and Bellezza (1989) uses the observed number of identical incorrect responses between two examinees. This analysis conditions upon the total number of errors in common between any two students, but it does not recognize that items vary in difficulty. The other difficulty with the analysis is that the observed number of identical incorrect responses depends upon the observed number of identical correct responses, which is a random variable and is not modeled in the analysis. Thus, this example supports Angoff’s position that the distribution of similarity statistics is not easily modeled theoretically. Building upon partially successful attempts by earlier researchers, distributions of more recently published similarity statistics (i.e., GBT, S-Check, and M4) appear to be well approximated using generalized binomial and trinomial distributions.

As discussed in the preceding paragraphs, some researchers have developed similarity statistics under the assumption that items are equally difficult or that wrong answers are equally attractive. For example, this assumption was made explicitly by Bellezza and Bellezza (1989) and implicitly by Angoff (1974). In fact, the author has attempted to develop computationally efficient algorithms for similarity statistics using the assumption that items are equally difficult (Maynes, 2013). He reported that the assumption results in overdetecting pairs of similar test responses (i.e., Type I error is inflated) because the expected value of the number of matching responses is underestimated (see Maynes, 2013 for mathematical details). The assumption that all items are equally difficult or that all wrong answers are equally attractive, while appealing and appearing to be trivial, is not supportable in practice. Because of this, the test response data should *not* be pooled across items to estimate matching probabilities. Doing so will result in approximating distributions that are biased toward the lower tail, which will spuriously raise false positive rates (i.e., the tail probabilities will be smaller than the actual probabilities).

Region of Permissible Values for Similarity Statistics

Even though the similarity statistics described in this chapter use examinee performance and item difficulty to estimate the statistical distributions, they ignore the fact that the distributions are bounded (especially when the normal approximation is used, as recommended by Wesolowsky). Maynes (2013) used the term “region of permissible values” to describe how the distributions of similarity statistics are bounded.

After conditioning upon test-taker performance (i.e., the number of correct answers), the distribution of the number of identical correct and incorrect responses is confined to a region of permissible values (Maynes, 2014). Values outside of this region are impossible. For example, once it is known that both test takers answered every question correctly, even one identical incorrect response would be impossible. Holland (1996) discusses the relationship between the number of matching (identical) incorrect answers and the number of answers where the two test takers disagreed. He emphasized that once the raw test scores are known (i.e., the total number of items answered correctly by each test taker), the *marginal totals* are fixed and constrain the possibilities for the two-by-two table of agreement between items answered correctly and incorrectly. An example from a 60-item test is shown in Table 3.1.

The marginal totals in Table 3.1 are provided in bold font because they are fixed. Given the data in Table 3.1, the greatest number of questions on which both Test Takers T_1 and T_2 could answer correctly is 42, and the lowest number is 27. In other words, the test takers MUST answer at least 27 of the same questions correctly and CAN-NOT answer more than 42 of the same questions correctly. Conversely, if 42 questions are answered correctly by both test takers, it MUST be the case that they answered the same 12 questions incorrectly and they disagreed upon the answers for the three remaining questions.

Table 3.1 Example Agreement Between Test Takers T_1 and T_2 with Scores of 42 and 45

T_1 / T_2	Correct	Incorrect	Total
Correct	42 to 27	0 to 15	42
Incorrect	3 to 18	15 to 0	18
Total	45	15	60

Table 3.2 Agreement Between Test Takers T_1 and T_2 With Scores of Y_1 and Y_2

T_1 / T_2	Correct	Incorrect	Total
Correct	$N_{11} = R$	$N_{12} = Y_1 - R$	Y_1
Incorrect	$N_{21} = Y_2 - R$	$N_{22} = N + R - (Y_1 + Y_2)$	$N - Y_1$
Total	Y_2	$N - Y_2$	N

Given the total number of questions (N) and the number of correct answers for the two test takers (Y_1 and Y_2), all of the cell counts in the two-by-two table of agreement will be determined when one other count has been established. For convenience sake, it is suitable to use the number of correctly answered questions (R) shared by the two test takers for this quantity. These quantities are shown in Table 3.2.

In Table 3.2, R (the number of identical correct answers) cannot exceed the minimum value of Y_1 and Y_2 and not less than the maximum value of 0 and $(Y_1 + Y_2 - N)$. Likewise, the value of N_{22} (the maximum number of identical incorrect responses) must be between the values of 0 and $N - \max(Y_1, Y_2)$. If there is only one correct answer for each item, R is the number of identical correct answers. If there is only one incorrect answer for each item (i.e., True/False question), $N + R - (Y_1 + Y_2)$ is the number of identical incorrect answers.

In summary, the following relationships define the region of permissible values:

1. If each question has only one correct answer and one incorrect answers (e.g., True/False), the region of permissible values is defined on an interval between $\max(0, N - Y_1 - Y_2)$ and $N + \min(Y_1, Y_2) - \max(Y_1, Y_2)$.
2. If each question has only one correct answer and multiple incorrect answers (e.g., typical multiple-choice question), the region of permissible values is defined by a triangular area with the number of identical correct answers, R , lying between $\max(0, Y_1 + Y_2 - N)$ and $\min(Y_1, Y_2)$, and with the number of identical incorrect answers lying between 0 and the value $R - \max(0, Y_1 + Y_2 - N)$. The total number of identical answers is defined on the interval between $\max(0, Y_1 + Y_2 - N)$ and $N + \min(Y_1, Y_2) - \max(Y_1, Y_2)$.
3. If each question has multiple correct answers and multiple incorrect answers (e.g., a math problem where several answer variants are correct), the region of permissible values is defined by a rectangular area with the number of identical correct answers, R , lying between 0 and $\min(Y_1, Y_2)$, and with the number of identical incorrect answers lying between 0 and the value $\min(Y_1, Y_2) - \max(0, Y_1 + Y_2 - N)$. The total number of identical answers is defined on the interval between 0 and $2[\min(Y_1, Y_2)] - \max(0, Y_1 + Y_2 - N)$.

While the above exercise in establishing the region of permissible values for the number of identical correct responses, R , and the number of identical incorrect responses, W , may seem trivial, it is of critical importance. None of the distributions of similarity statistics published to date have accounted for this restriction. Instead, mathematical formulae have been used to compute probabilities for tail values that are impossible to observe. Thus, the computations published for the distributions of GBT, S-Check, and M4 only approximate actual probabilities.

To understand the approximation errors that result from ignoring the region of permissible values, it is useful to overlay the region of permissible values onto the bivariate

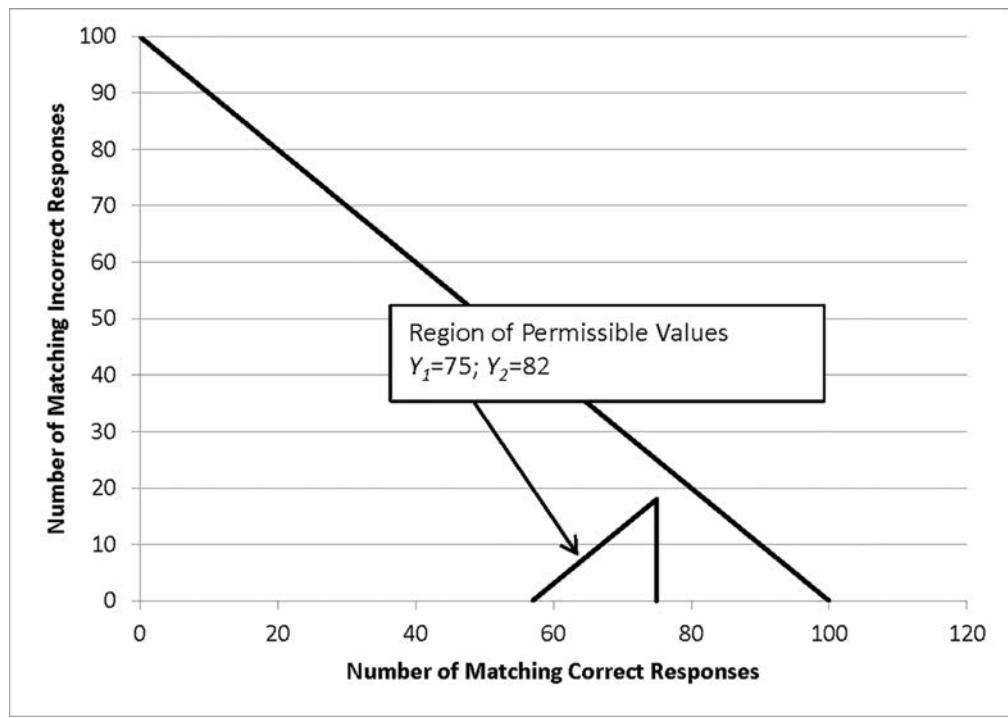


Figure 3.1 Illustration of region of permissible values; 100 items

sample space of matching correct and matching incorrect answers. This has been done in Figure 3.1.

In Figure 3.1, the upper line that begins at the point (0, 100) and ends at (100, 0) represents the boundary of the number matches that could possibly be observed. It defines the probability space that is modeled by the GBT and M4 statistics. Probabilities are explicitly or implicitly computed for every single one of the points within this triangle. The region of permissible values for two test takers with scores of 75 and 82 is shown in the smaller triangle with vertices at (75, 0), (82, 0), and (75, 18). Each of the similarity statistics computes and adds up probability values that are impossible to observe, being outside the region of permissible values. The question remains whether this neglect is important. For the generalized binomial statistic, GBT, probabilities will be assigned to impossible lower tail values. Probabilities will also be assigned to impossible upper tail values. In general, computing probabilities for impossible upper tail values will result in computing tail probabilities that are too large. Thus, the statistical procedures will tend to overstate Type I errors. The actual Type I error rate will be lower than the reported probability. This results in loss of statistical power. This effect is more pronounced as the difference between Y_1 and Y_2 increases. A similar analysis applies to M4, with the exception that the probabilities are summed using a curved contour, which results in closer approximations to actual probabilities. Computed tail probabilities will still be larger than the actual probabilities (this result is seen in the simulation of M4 Type I errors in this chapter).

The S-Check statistic uses a normal approximation which has an infinite tail. Thus, tail probabilities computed by this statistic are likely to be larger than the same probabilities computed by the GBT and M4 statistics, meaning this statistic will lose more power than the other two statistics.

THE M4 SIMILARITY STATISTIC

This section discusses the M4 similarity statistic because it appears to provide the best approximation for computing the probability of observed similarity (i.e., it is least impacted by noncompliance with the region of permissible values) and because it is more appealing to use two separate pieces of evidence (i.e., the number of identical correct and identical incorrect responses) that have varying power than to combine them and possibly dilute the strength of the evidence.

The M4 similarity statistic is a bivariate statistic. It consists of the number of identical correct responses and the number of identical incorrect responses. It has been generally acknowledged that incorrect matching answers provide more evidence of nonindependent test taking than correct matching answers, because these are low-probability events. Thus, a bivariate statistic should be able to take advantage of this fact statistically.

The M4 similarity statistic is more appealing than the GBT and S-Check statistics because it allows the incorrect matching answers and the correct matching answers to be evaluated jointly, but separately.

Probability Density Function and Tail Probabilities for the M4 Similarity Statistic

The probability density function of the M4 similarity statistic may be approximated by assuming statistical and local independence of matching responses. Using these assumptions, the statistic follows a generalized trinomial distribution. This distribution is a special case of the generalized multinomial distribution. It does not have a closed form, but its generating function (Feller, 1968) is

$$G(x, y) = \prod_i (r_i + p_i x + q_i y), \quad (\text{Equation 1})$$

where p_i is the probability of a matching correct response for item i, q_i is the probability of a matching incorrect response for item i, r_i is the probability of a nonmatching response for item i, x is the count of observed matching correct responses, and y is the count of observed matching incorrect responses. The reader should note that $p_i + q_i + r_i = 1$.

Using the generating function of the generalized trinomial distribution (Equation 1), the joint probability distribution of x and y may be computed using a recurrence relation (Tucker, 1980, see also Graham, Knuth, & Patashnik, 1994). For a pair of responses, there are three possibilities: a matching correct, a matching incorrect, and a nonmatching response. Let these possibilities be summarized using three triplets: (1,0,0), (0,1,0), and (0,0,1), where the first value of the triplet is a binary value indicating whether there was a matching correct answer, the second value indicates whether there was a matching incorrect answer, and the third value indicates whether there was a nonmatching number. Because the third value is equal to one minus the sum of the other two values, it can be discarded. Hence, the trinomial distribution can be formed using bivariate pairs of (1,0), (0,1), and (0,0).

Using the above notation, the trinomial distribution for M4 can now be expressed mathematically. The joint probability distribution for M4 is written as

$$\begin{aligned} T_{k+1}(x, y) &= p_{k+1}(1,0)T_k(x-1, y) \\ &\quad + q_{k+1}(0,1)T_k(x, y-1) \\ &\quad + (1 - p_{k+1}(1,0) - q_{k+1}(0,1))T_k(x, y), \text{ with boundary condition} \end{aligned} \quad (\text{Equation 2})$$

$T_0(0,0) = 1$ and $T_0(x, y) = 0 \forall (x, y) \neq (0,0)$,

where the values of $T(x,y)$ are computed successively for the matches of each item response pair represented by the subscript $k+1$, x is the number of matching correct responses, and y is the number of matching incorrect responses. The value of $k+1$ begins with 1 and ends with n (the number of items answered by both test takers). In Equation 2, the value of r_{k+1} is implicitly computed by subtraction (i.e., $1-p_{k+1}(1,0)-q_{k+1}(0,1)$). The bivariate pairs of (1,0), and (0,1) are explicitly shown in Equation 2 to emphasize that the probabilities are associated with matching correct, matching incorrect and nonmatching responses. When the values of p_i and q_i are constant, the generalized trinomial distribution becomes the trinomial distribution.

Recurrence equations (such as Equation 2) can be a bit difficult to understand in symbolic notation. An illustration shows how the recurrence is computed.

Table 3.3 illustrates four subtables. The first subtable corresponds to the initial condition, when with probability one there are no matching correct and no matching incorrect responses. The succeeding subtables correspond to the joint probability distribution that results after adding an item to the recurrence. For the first item, the probability of a correct match is 0.30 and the probability of an incorrect match is 0.20, likewise the two corresponding probabilities for the second item are 0.35 and 0.15, and for the third item they are 0.45 and 0.25. The rows in each subtable correspond to the number of correct matches. The columns correspond to the number of incorrect matches. The subtables are triangular shaped because the number of correct identical matches added to the number of incorrect identical matches cannot exceed the number of items.

Equation 2 specifies how to compute the values for each cell in the table. For example, the value of the cell for $T_3(1,1)$ is computed by substituting p_3 and q_3 into the formula to obtain $0.45 \times T_2(0,1) + 0.25 \times T_2(1,0) + 0.30 \times T_2(1,1)$.

Table 3.3 Recurrence Pattern for the Generalized Trinomial Distribution

T ₀ : 0/0 (initial)	0	
	0	1

T ₁ : 0.3/0.2 (first item)	0	1	
	0	0.5	0.2
	1	0.3	

T ₂ : 0.35/0.15 (second item)	0	1	2	
	0	0.25	0.175	0.03
	1	0.325	0.115	
	2	0.105		

T ₃ : 0.45/0.25 (third item)	0	1	2	3
	0	0.075	0.115	0.05275
	1	0.21	0.1945	0.04225
	2	0.17775	0.078	
	3	0.04725		

It is important to realize that the trinomial distribution has three tails, as illustrated in Table 3.3 and Figure 3.1. As a result, standard probability computations that are used for univariate distributions are not applicable. There is no obvious direction for computing the tail probability because there is no upper tail and no lower tail. Because there is not an obvious upper tail, the desired tail probability must be computed by using a subordering principle as recommended by Barnett & Lewis (1994).

The subordering principle is implemented using a two-step procedure. First, an “upper” probability is computed for each point (x,y) in the M4 distribution. This is done by adding the probabilities for all bivariate points (t,v) where t is greater than or equal to x or v is greater than or equal to y . This quantity is named $D_{x,y}$. Second, the subordering principle stipulates that the desired probability associated with the point (x,y) is the sum of all values of $T(j,k)$ where $D_{j,k} \geq D_{x,y}$. The subordering principle for the probability computation is illustrated in Figure 3.2.

In Figure 3.2, the diamond provides the location of the expected value of the distribution and the square provides the observed value for a pair of extremely similar test instances. Step one denotes addition of the bivariate probabilities from the observed data to point to the triangular boundary in the direction of upward and to the right. If the probability computation stops with step one, *the tail probability will be an underestimate of the actual probability* (i.e., the probability of rejection will be inflated) because the subordering principle has not yet been applied completely. The values from step one are used in defining the subordering principle. The data points in the bivariate distribution are ordered using the values computed in step one. The bivariate probabilities associated with the ordered points are summed to compute the tail probability. This is represented as step two in Figure 3.2.

Thus, the probability of the observed value is computed by using a directional ordering relation that defines the direction of extremeness as being upwards (a greater number of identical incorrect answers) and to the right (a greater number of identical correct answers) of the observation. After having computed the probability using the directional ordering relation, the definition of extremeness is a direct interpretation of

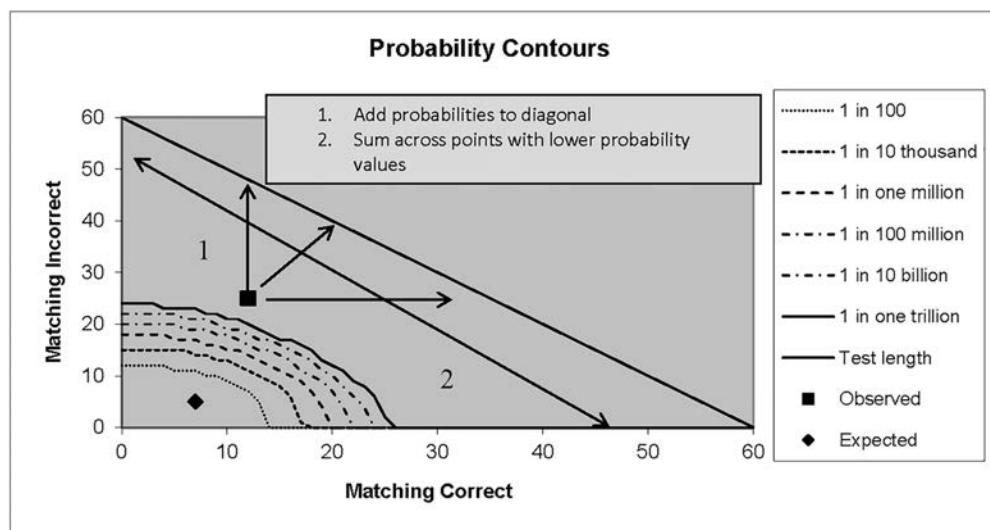


Figure 3.2 Illustration of the M4 Probability Computation

the probability. Observations with smaller probabilities are more extreme than other observations.

Analysis of Licensure Data Set

The M4 statistic can be used both to detect potential cases of nonindependent test taking and to confirm or reject propositions that particular individuals may have committed test fraud. Detection of potential test fraud involves data mining and examining all pairs of interest in the data set. Because a large number of pairs will be evaluated, the likelihood of committing a Type I error (i.e., incorrectly stating that a test instance was not taken independently) will be high unless an appropriate error control is used. The multiple comparisons used in the analysis of the licensure data set was based on the maximum order statistic (Maynes, 2009). This correction is written as

$$I_{max,s} = -\log_{10}(1-(1-p_{min})^{(N_s-1)/2}), \quad (\text{Equation 3})$$

where $I_{max,s}$ is the index for the examinee computed using N_s , the number of test takers that were compared, and p_{min} , the smallest observed probability value. Additionally, because M4 is computationally intense, the licensure data set was split in two subsets, each subset corresponding to one of the forms. As was done with the simulation, the responses to the pretest items were discarded. There were a total of 2,687,976 pairwise comparisons performed.

The detection threshold was set using the procedure described in the “Discussion of Exploratory and Confirmatory Analyses” section. That procedure stated that the probability value should be adjusted using $(N_s-1)/2$ (N_s was set at 1,636 and 1,644 for Form 1 and Form 2, respectively). This was done using Equation 3. It further stated that after comparing the test response for each test taker, the probability value should then be adjusted using N (3,280). The second adjustment means that the detection threshold was set at an index value of 4.82, after adjusting for the number of test takers per form. Detection using this threshold maintains a low false positive rate of 0.05 while providing a good possibility of detecting true positives. The Bonferroni calculation for the second adjustment was $p = 0.05/2687976$, which is approximately 1 in 53 million. The simulation results in Table 3.5 in this chapter indicate that a significant number of nonindependently answered items (e.g., perhaps 30% to 50%) would be necessary in order for a pair to be detected using this procedure. Table 3.4 lists the pairs that were detected in the analysis of the licensure data.

In Table 3.4, the Flagged column indicates whether the examinee is a known cheater. The next three columns indicate the Examinee’s country, state in which the examinee applied for licensure, and the school or institution where the examinee was trained. The Test Site indicates where the test was administered. The Pass/Fail Outcome indicates whether the examinee had a passing score on the exam. And the M4 Similarity Index is the observed probability for the pair adjusted using Equation 3.

From among 2,687,976 pairs that were compared, M4 detected seven pairs. If this procedure were performed on this many tests which were taken independently, on average only one pair would be detected every 20 analyses. In other words, the expected number of detected pairs using the Bonferroni adjustment was 0.05.

The reader will notice in Table 3.4 that some individuals were detected multiple times. Detections that involve more than two individuals are known as clusters and are discussed by Wollack and Maynes (this volume).

Table 3.4 Detections of Potentially Nonindependent Pairs in the Licensure Data

Examinee ID	Flagged	Country	State	School	Exam Form	Test Site	Site State	Pass/Fail Outcome	Raw Score	M4 Similarity Index
e100624	1	India	28	5530	Form1	2305	28	1	140	9.5
e100505	1	India	28	8152	Form1	2305	28	1	132	9.5
e100505	1	India	28	8152	Form1	2305	28	1	132	7.7
e100498	1	India	42	8119	Form1	5880	42	1	136	7.7
e100505	1	India	28	8152	Form1	2305	28	1	132	5.9
e100452	1	India	28	8172	Form1	2305	28	1	128	5.9
e100624	1	India	28	5530	Form1	2305	28	1	140	5.5
e100452	1	India	28	8172	Form1	2305	28	1	128	5.5
e100226	1	India	42	8155	Form1	5303	54	0	112	5.2
e100191	1	India	42	5530	Form1	2305	28	0	107	5.2
e100505	1	India	28	8152	Form1	2305	28	1	132	5.0
e100494	1	India	42	8198	Form1	5302	54	1	131	5.0
e200294	0	India	42	8198	Form2	2305	28	0	113	6.5
e200448	1	India	28	8198	Form2	2305	28	1	126	6.5

Presentation of Results from M4 Similarity Analysis

To plot and evaluate M4, two statistically dependent values are computed when comparing two test instances: (1) the number of identical correct responses and (2) the number of identical incorrect responses. These values may be plotted using a pie chart with three slices. Examples of these data are shown in Figure 3.3, using the first pair listed in Table 3.4.

The panel on the left of Figure 3.3 provides the counts of observed similarities between the two tests, and the panel on the right of Figure 3.3 provides the expected agreement on the tests. There were 170 scored questions analyzed in this comparison.

As a means of comparison, two other test takers, e100553 and e100654, with exactly the same scores as the pair illustrated in Figure 3.3, were selected from the licensure data. For this new pair, the M4 similarity index was equal to 0.3 (the median value of the distribution). These data are shown in Figure 3.4.

The format and labels of the data in Figure 3.4 are the same as those in Figure 3.3. The observed numbers and expected numbers of agreement in Figure 3.4 are nearly equal, which is expected.

The counts of observed and expected agreement may be more readily visualized when the selected responses of the common items are aligned. For the sake of simplicity in the presentation and exam security, the actual item responses are not shown. Instead, each item response is shown as being the same correct response, the same incorrect response, or differing responses between the two exams. The data for Figure 3.3 and Figure 3.4 are shown in Figure 3.5.

There are two panels in Figure 3.5. The upper panel depicts the alignment data for the 170 items for the two test takers whose data were shown in Figure 3.3 (i.e., the extremely similar pair—index value of 12.5). The lower panel depicts the alignment data for the same items for the two test takers whose data were shown in Figure 3.4

Extreme Similarity – Index = 12.5

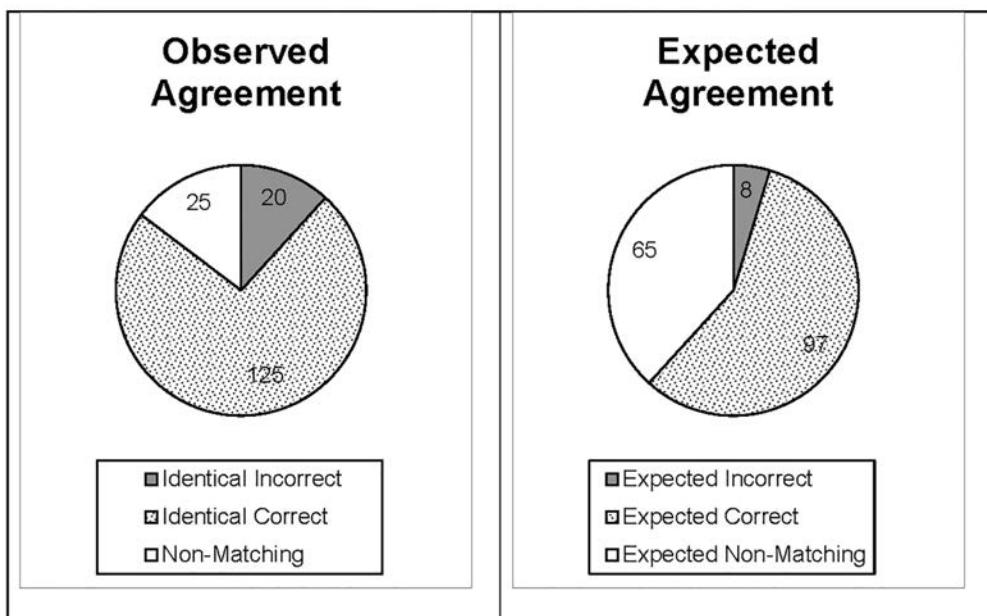


Figure 3.3 Illustration of the M4 Similarity Statistic: Extreme Similarity Index = 12.5

Median Similarity – Index = 0.3

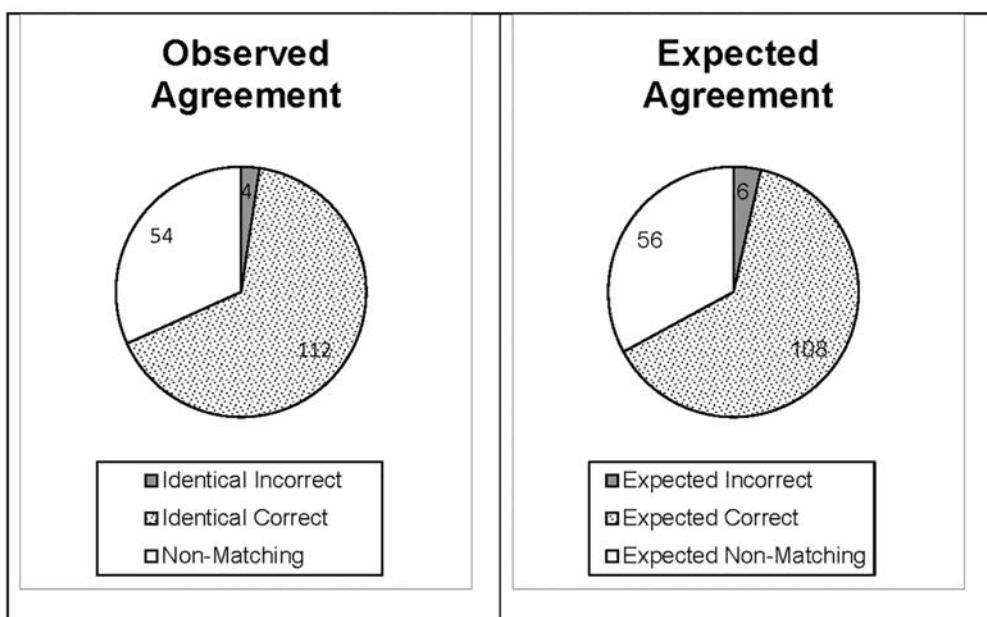


Figure 3.4 Illustration of the M4 Similarity Statistic: Median Similarity Index = 0.3



Figure 3.5 Illustration of Aligned Responses

(i.e., the typical similarity pair with an index value of 0.3). As was shown in Figures 3.2 and 3.3, the numbers of agreed upon incorrect and correct responses for the extremely similar pair were much higher than for the typical similarity pair.

The extremity of an observed value of M4 may be shown using a contour plot of the bivariate probability distribution for the number of identical correct and incorrect responses. The contour plot for the extremely similar pair (i.e., the test takers from Figure 3.1 with an index of 12.5) is shown as Figure 3.6.

Figure 3.6 illustrates the region of the probability distribution for the bivariate distribution of the matching correct and matching incorrect response counts. The observed data from Figure 3.3 (i.e., 20 matching incorrect and 125 matching correct responses) have been plotted using a large square, and the expected values of the distribution (i.e., 8 matching incorrect and 97 matching correct responses) have been plotted using a diamond. The curved lines (labeled 1 in 100, 1 in 10 thousand, etc.) between the expected and observed values represent diminishing probability contours. Each contour line is drawn at a power of 100. The contour with the lowest probability represents a probability level of 1 chance in 10^{12} , or one in a trillion.

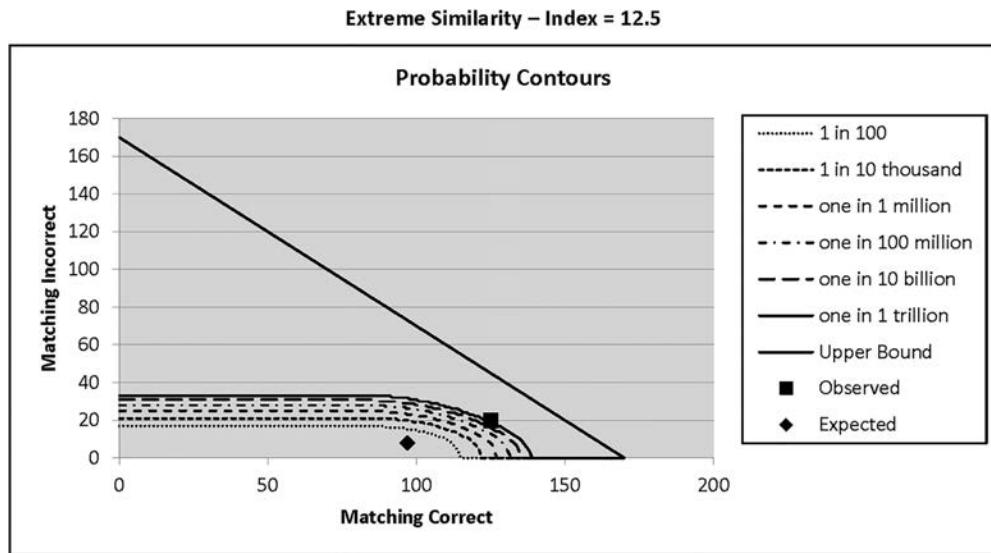


Figure 3.6 Contour Plot of the M4 Similarity Statistic: Extreme Similarity Index = 12.5

The direction of probability computation is chosen to maximize evidence for the alternative hypothesis of nonindependent test taking when the data are extreme and due to collusion. For example, the point (0, 0) is very far from the expected value of the distribution, but it does not provide evidence that the tests are similar. The upper bound is the limit of the distribution where the number of identical correct and identical incorrect answers equals the number of items (meaning that the number of nonmatching answers is zero, because the total number of items is the sum of identical correct, identical incorrect, and nonmatching responses).

Analysis of Type I and Type II Errors for M4 Similarity

A simulation was conducted to evaluate Type I and Type II error rates for M4. The simulation was performed using the licensure data set provided to chapter authors in this book. The licensure data set was prepared by removing the responses for the pre-test items, leaving 170 scored items on each form. Even though the M4 statistic can be used to create clusters of nonindependently taken exams, in this section the simulation was conducted using pairs. Parameters for the NRM were estimated using the provided licensure data. These parameters were then used to simulate item responses.

Type I errors result from incorrectly rejecting the null hypotheses. Type II errors result from failure to reject the null hypotheses when the null hypotheses are false. Type I errors will potentially produce inappropriate score invalidations. Type II errors will result in acceptance of inappropriate test scores. Thus, the practitioner must balance the costs of Type II errors against the costs of Type I errors and the benefits of correctly detecting nonindependent test taking against the benefits of correctly determining that the test was taken appropriately. The primary mode by which these costs and benefits may be balanced is through selection of the flagging threshold for the similarity statistic. Conservative thresholds reduce Type I errors at the expense of increased Type II errors. Statistical decision theory can provide guidance for selecting the threshold, but ultimately determination of the threshold is a policy decision.

Simulation Methodology

Twelve data sets were generated using simulation under different levels of dependence, varying from 0% copying to 55% copying in increments of 5%. Each of the 12 data sets was generated in the following manner:

1. Each test response vector in the live data set was selected and used as a “reference” vector 31 times, thereby allowing (a) source data to be actual test data and (b) a total of over 100,000 pairs of examinees to be simulated (i.e., $3,280 \times 31$ equals 101,680).
2. The value of θ was computed for the reference vector. Two values of θ , θ_1 and θ_2 were sampled from a normal distribution using θ as the mean and a variance of one.
3. Using a prespecified proportion of copying (i.e., varying from 0% to 55%), items were randomly sampled without replacement from the reference vector and selected for copying.
4. Two item response vectors (IRV), ν_1 and ν_2 , were then generated using θ_1 and θ_2 , respectively. Item responses were copied from the reference vector of those items that were selected for copying and were generated using θ_1 and θ_2 from the NRM for the remaining items. Even though a source-copier model was not simulated, this is functionally equivalent to simulating the entire IRV for two copiers and then changing responses to match the source (i.e., the reference vector) for the appropriate proportion of copied items.

The above approach allows for simulating the standard source-copier setup. It also allows for simulating general collusion where the “copied” items become “disclosed” or “shared” items. It also may be extended to simulate other collusion scenarios. However, in the present case, only simulated pairs of IRV’s are needed.

Simulation Results—Null Condition

The null condition corresponds to the 0% copying level. This is the condition when tests are taken independently and is the customary null hypothesis for the M4 statistic. When M4 is used for detection and/or confirmation of tests taken nonindependently, the upper tail value is evaluated. If the NRM is appropriate for the test response data, the index value of M4 follows an exponential distribution. The index value is mathematically equivalent to the upper tail probability ($p = 10^{-\text{index}}$). Thus, an index value of 1.301 corresponds with an upper tail probability of 0.05. Because each pair was examined only once, the Bonferroni correction was not used in the simulations. Table 3.5 summarizes the tail probabilities estimated from the simulated condition of 0% copying.

Table 3.5 Type I Error Rates When Tests Are Taken Independently

Index	Theoretical Rate	Number of Detections	Observed Rate
1.301	0.05	3,155	0.031
2.0	0.01	520	0.0051
2.301	0.005	268	0.0026
3.0	0.001	44	0.00043
3.301	0.0005	17	0.00017
4.0	0.0001	1	0.00001

Table 3.6 Type II Error Rates for Copying Levels From 5% to 55%

Copying Condition	Detection Thresholds Specified as the M4 Similarity Index Value					
	1.301	2.0	2.301	3.0	3.301	4.0
5%	0.87891	0.96790	0.98213	0.99547	0.99752	0.99931
10%	0.65582	0.86510	0.91226	0.96929	0.98140	0.99445
15%	0.38748	0.65773	0.74347	0.87983	0.91487	0.96408
20%	0.15572	0.37209	0.46875	0.66862	0.73634	0.85555
25%	0.05381	0.16902	0.23685	0.41233	0.49002	0.65446
30%	0.01296	0.05371	0.08386	0.18462	0.23885	0.37953
35%	0.00306	0.01681	0.02887	0.07522	0.10291	0.19037
40%	0.00077	0.00400	0.00747	0.02237	0.03319	0.07023
45%	0.00013	0.00095	0.00199	0.00690	0.01061	0.02540
50%	0.00004	0.00022	0.00033	0.00160	0.00255	0.00662
55%	0.00000	0.00001	0.00010	0.00044	0.00068	0.00218

Table 3.5 shows that the simulation probabilities for the M4 Similarity statistic are close to, but less than, the theoretical probabilities, showing that M4 has good control of the nominal Type I error rate. As mentioned in the discussion of the region of permissible values, the theoretical rates for M4 are greater than observed rates. The difference is due to the way in which M4 approximates the actual probability value.

Simulation Results—Copying Conditions

Table 3.6 provides the false negative or Type II error rates for selected index values of M4.

As expected, the Type II error rate for M4 decreases as the copying rate increases or, conversely, statistical power² increases as the copying rate increases. At copying levels of 50% or greater, the statistic has very high power. Additionally, the Type II error rate increases as the detection threshold increases.

SUMMARY

This chapter has provided an exposition for a method of detecting nonindependent test taking by comparing pairs of test instances using similarity statistics. A simulation was performed that showed (1) the Type I error rate is not inflated using the M4 similarity statistic, and (2) a significant number of identical responses are needed before a pair of test takers will be detected as potentially not having taken their tests independently. Future papers are needed to describe and explore clustering approaches, computing cluster-based probability evidence for nonindependence, or evaluating performance changes that may have resulted through the behavior that produced the cluster of similar test instances. The problem of cluster extraction is very important, and Wollack and Maynes (this volume) explore the accuracy of extraction for a simple clustering algorithm.

NOTES

1. The approximation of the Bonferroni adjustment is best when p is small. When p is large, the distribution of the maximum order statistic provides a better approximation (Maynes, 2009).
2. Statistical power is the ability to detect true positives. It is equal to 1 minus the Type II error rate.

REFERENCES

- Allen, J. (2014). Relationships of examinee pair characteristics and item response similarity. In N. M. Kingston and A. K. Clark (Eds.) *Test fraud: Statistical detection and methodology* (pp. 23–37). Routledge: New York, NY.
- Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, 69, 44–49.
- Barnett, V. & Lewis T. (1994). *Outliers in statistical data*, 3rd Edition, pp. 269–270. John Wiley and Sons: Chichester, UK and New York, NY.
- Bellezza, F. S. & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology*, 16, 151–155.
- Bellezza, F. S. & Bellezza, S. F. (1995). Detection of copying on multiple choice test: An update. *Teaching of Psychology*, 22, 180–182.
- Belov, D. I. (2013). Detection of test collusion via Kullback–Leibler divergence. *Journal of Educational Measurement*, 50(2), 141–163.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 46, 443–459.
- Feller, W. (1968). *An introduction to probability theory and its applications*. Volume I, 3rd Edition (Revised Printing). John Wiley & Sons, Inc.: New York, London, Sydney. p. 279.
- Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, 6, 152–165.
- Graham, R. L., Knuth, D. E., & Patashnik, O. (1994). *Concrete mathematics*, 2nd Edition. Addison-Wesley Publishing Company: Reading, MA. Chapter 7.
- Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying*. ACT Research Report Series. 87–15. September 1987.
- Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index*. ETS Program Statistics Research. Technical Report No. 96–4.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates: Hillsdale, NJ.
- Maynes, D. D. (2009, April). *Combining statistical evidence for increased power in detecting cheating*. Presented at the annual conference of the National Council on Measurement in Education, San Diego, CA.
- Maynes, D. D. (2013, October). *A probability model for the study of similarities between test response vectors*. Presented at the Conference of Statistical Detection of Potential Test Fraud, Madison, WI.
- Maynes, D. D. (2014). Detection of non-independent test taking by similarity analysis. In N. M. Kingston and A. K. Clark (Eds.) *Test fraud: Statistical detection and methodology* (pp. 53–82). Routledge: New York, NY.
- Tucker, A. (1980). *Applied combinatorics* (pp. 111–122). John Wiley & Sons, Inc.: New York, Brisbane, Chichester, Toronto.
- van der Linden, W. J. & Sotaridona, L. S. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, 31(3), 283–304.
- Wesolowsky, G. O. (2000). Detection excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics*, 27(7), 909–921.
- Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, 21(4), 307–320.
- Wollack, J. A., Cohen, A. S., & Serlin, R. C. (2001). Defining error rates and power for detecting answer copying. *Applied Psychological Measurement*, 25(4), 385–404.
- Wollack, J. A., & Maynes, D. D. (2011, February). *Data forensics: What works, what doesn't*. Presentation at the annual conference for the Association of Test Publishers, Phoenix, AZ.
- Zhang, Y., Searcy, C. A., & Horn, L. (2011, April). *Mapping clusters of aberrant patterns of item responses*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

4

IDENTIFYING AND INVESTIGATING ABERRANT RESPONSES USING PSYCHOMETRICS-BASED AND MACHINE LEARNING-BASED APPROACHES¹

Doyoung Kim, Ada Woo, and Phil Dickison

A test is high-stakes if its results are used to make important decisions about test takers. In general, test takers take high-stakes tests to receive benefits such as a license to practice an occupation or to avoid punishments (e.g., being denied a diploma, not being permitted to drive a car). As more high-stakes tests are administered to more people for diverse reasons, the validity of test scores becomes important. It is imperative that any person or organization who uses high-stakes test scores to make such important decisions about test takers be confident that the scores from the test takers are reliable and provide valid indications of the test takers' abilities on the construct being measured by the test. In the recent International Testing Committee (ITC) guidelines on quality control in scoring, test analysis, and reporting of test scores (International Test Commission, 2013), it is recommended that aberrant or unexpected response patterns (e.g., missing easy items while answering difficult ones correctly) should be monitored routinely through statistical techniques for detecting invalid test scores. Because the aberrant item response patterns have a negative impact on the validity of test scores, it is important to identify and investigate these item response patterns prior to drawing any conclusion using the test scores. As computer-based testing has become popular, test takers' response times are readily available, and having item responses accessible can facilitate detection of invalid test scores. These response time data are not only important themselves in that aberrant response time patterns can be identified by fitting statistical models to the response time data but are also valuable to bring a new perspective to investigating aberrance when using together with the response pattern data. Furthermore, some of the auxiliary information collected with item response and response time data, such as test takers' demographic variables and testing center environment, could also inform any investigation into the individual examinees whose responses are aberrant.

VALIDITY TRIANGLE

Aberrant behavior during testing is not a new phenomenon. Examples of aberrant behavior by test takers have been reported as early as 165 BC, when applicants for the Chinese Civil Service examinations devised methods to gain advantage over other applicants taking the examination (Callahan, 2004). The earliest reported attempt at a systematic, evidenced-based approach at detecting these cheating behaviors was focused on probabilistic models identifying collusion and/or answer copying (Bird, 1927). The initial work by Bird sparked decades of investigation by testing professionals into refining and improving evidenced-based models focused on identifying individual cheaters (Haney & Clarke, 2006). This research has provided significant information to help test developers, test administrators, and psychometricians to improve processes and procedures to deter activities of test takers to gain an unfair advantage when interacting with a test.

Although the efforts of testing professionals to identify cheating behavior are laudable, they have not been extremely effective in slowing the growth of cheating by individuals (Pérez-peña, 2012). The reasons for this growth are multifaceted. The lines of inquiry generally focus on moral, ethical, and societal factors (Callahan, 2004) and thus testing professionals' efforts to prove aberrant test behavior are conflated with significant societal and ethical issues. The political, social, and legal implications associated with this conflagration has resulted in a paradigm where data suggesting a high probability of aberrant test behavior is directed at proving moral intent, ethical malfeasance, and degree of harm to the individual participating in the aberrant behavior. As early as the 1970s, testing professionals have identified the difficulty of using statistics to identify individuals as cheaters when framed in a paradigm of cultural constructs (Dwyer & Hecht, 1994; Frary, Tideman, & Watts, 1977; van der Linden & Jeon, 2012; Wainer, 2014).

The use of data to identify statistically improbable test behaviors is defensible. However, generalizing the results to a testing individual or population to prove cheating is problematic. Inferences about cheating must also include evidence demonstrating motive, intent, opportunity, damages, and means. Inferential statistics are based on a priori research questions that guide the methodology and provide a link between the observed statistic and the population of interest. Using this argument, the stated goal of identifying aberrant test behaviors should be related to validity of the scores not identifying cheaters, thus moving the argument from a complex ethical/cultural paradigm to a test validity paradigm.

The validity triangle (see Figure 4.1) demonstrates the importance and equality of the relationship between psychometrics, content, and security to the validity argument of any test. In addition, it demonstrates how statistical inferences related to aberrant test behavior are directly connected to test scores and subsequently test validity. This allows testing professionals to disconnect their probabilistic arguments from the cultural aspects surrounding the actions of an individual test taker and focus them solely on the veracity of the test scores.

Testing professionals have been increasingly asked to evaluate the veracity of the test scores obtained from testing (Meijer, Tendeiro, & Wanders, 2014). Aberrant test behaviors are major threats to the veracity of the test scores. In the following section, two types of aberrant test behaviors are defined: *aberrant response patterns* and *aberrant*

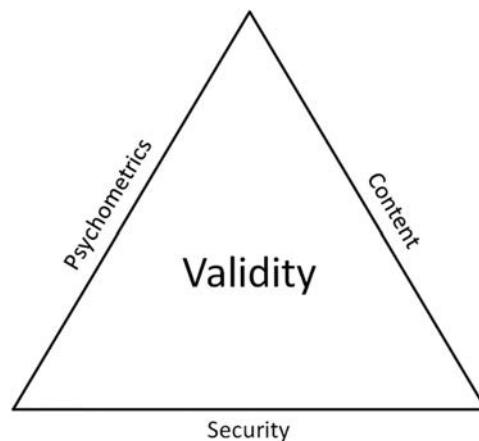


Figure 4.1 Validity Triangle

response time patterns. In addition, we explore potential underlying causes for these patterns. Subsequently introduced are how these patterns can be identified and how to investigate further the identified (or flagged) aberrant test takers.

OVERVIEW OF IDENTIFYING AND INVESTIGATING ABERRANT PATTERNS

Aberrant responses can be defined as unexpected responses that are not consistent with what a model predicts. Likewise, aberrant response times are deviations from expected response times. Researchers (Karabatsos, 2003; Meijer, 1996) have identified at least five factors that produce aberrant response patterns: random responding, cheating, creative responding, lucky guessing, and careless responding. Random responding occurs when test takers respond to multiple-choice items by randomly choosing an option for a set of items on the test. Cheating can be defined broadly to include any activity involving unacceptable means for accomplishing higher test scores (e.g., when an individual taker copies from another's test and a group-level behavior such as inappropriate coaching and test preparation). Creative responding happens when able test takers interpret relatively easy items in a creative way so that they incorrectly respond to those items. Lucky guessing is one of the outcomes of random responding and occurs when test takers guess the correct answers to some of items. Careless responding happens when test takers haphazardly respond incorrectly to some relatively easy items.

In addition to these factors, preknowledge of some of the items in a test or an item pool has been a focus of aberrant response research since item preknowledge is one of prevalent forms of cheating in computer-based testing [CBT] (Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014; McLeod, Lewis, & Thissen, 2003; Meijer & Sotaridona, 2006). When test takers have item preknowledge, their item response and time response patterns may deviate from expected (or model predicted) patterns. These are not exhaustive lists of the causes of aberrant response and time patterns, but they provide a good start for testing professionals when trying to understand the underlying causes of test aberrance.

Two approaches introduced in the following sections can provide useful tools for testing professionals. One is based on psychometrics and the other on machine

learning. Psychometrics-based methods help test professionals identify aberrant item response/time response patterns, and a machine learning-based approach can utilize unused variables (e.g., demographic variables and testing center environment) under psychometrics-based methods to uncover dominant characteristics associated with the aberrant patterns. Several person-fit indexes based on psychometrics are explained first in the next section, followed by response time models. One method based on machine learning—market basket analysis—is introduced. This method comes in handy, especially once test professionals detect or flag test takers who exhibit aberrant test behaviors. A data analysis section follows each of three method groups (person-fit indexes, response time models, and machine learning). Each data analysis section shows how aforementioned method(s) can be applied to the common licensure dataset that are used in several chapters in this volume.

PSYCHOMETRICS-BASED APPROACHES

Most of psychometrics-based approaches utilize either item responses or response times or both for identifying aberrant patterns. Person-fit analyses investigate whether the response patterns from (suspected) test takers deviate from the response patterns from the remaining group under a statistical model of interest (Rupp, 2013). Person-fit indexes are designed to identify persons with aberrant response patterns. Olson and Fremer (2013) published a report for the Council of Chief State School Officers (CCSSO) in which they advocated using, besides other methods, person-fit indexes to detect irregularities in test behavior. See Karabatsos (2003) and Meijer and Sijtsma (2001) for extensive reviews of person-fit indexes.

The response time (RT) is the time used by an examinee responding to an item in a test. It was difficult to collect RTs prior to the advent of Computer-Based Testing (CBT). The use of CBT allows test sponsors to reliably record RTs as a matter of routine. Readily available RT data have boosted research on RT models (see, e.g., Meijer & Sotaridona, 2006; Thissen, 1983; and van der Linden, 2006, 2007). RT models are useful for describing typical RT behavior, but in the same spirit as person-fit analyses, can also be used to identify individuals with unusual RT patterns that are significantly different from what is predicted under the model given their overall performance and rate of responding.

Person-Fit Indexes

Parametric Person-Fit Indexes

A number of parametric person-fit indexes have been developed to identify a person's response pattern that is not consistent with the response pattern expected from models based on IRT. Variants of the likelihood-based l_z index were most frequently investigated among published person-fit studies (Rupp, 2013). This popularity of the l_z index is because it is easy to calculate, and its asymptotic distribution follows a standard normal distribution. Hence, l_z values can be treated as z scores, and critical values on the unit normal can be used to flag persons with extreme l_z values. In addition, several studies (Armstrong, Stoumbos, Kung, & Shi, 2007; Drasgow, Levine, & McLaughlin, 1991; Li & Olejnik, 1997) reported that the l_z index is powerful to detect aberrant response patterns. In this section, we utilize two parametric person-fit indexes, the l_z index and its improved version, the l_z^* index, to detect aberrant responses. In addition

to l_z and l_z^* indexes, the performance of the Cumulative Sum (CUSUM) statistic is examined. CUSUM is different from commonly used person-fit statistics in that CUSUM provides information on aberrant patterns as they occur during a computer-based test rather than calculating a single value to represent the degree of aberrance for an entire response pattern for a given test taker. CUSUM monitors each response and signals when aberrant patterns occur, like an electrocardiogram (ECG or EKG) does to your heart.

l_z and l_z^* Indexes

The l_z index is the standardized version of the log-likelihood function l_0 (Drasgow, Levine, & Williams, 1985). A log-likelihood function is defined for a probability distribution. It is assumed that the probability distribution of the dichotomously scored response (X_{ij}) of item i from person j follows a Bernoulli distribution. The random variable, X_{ij} , takes only two values: 1 with success probability (P) and 0 with failure probability ($1 - P$). A dichotomous IRT model specifies the success probability for item i and person j .

IRT models for dichotomously scored items are often described by one of three widely used IRT models: the one-parameter logistic (1PL) model (a special case of which is the Rasch model), the 2PL model, and the 3PL model. The most general model among the three dichotomous IRT models, the 3PL model, describes the relationship between the probability of person j ($j = 1, \dots, N$) answering correctly to item i ($i = 1, \dots, n$) and the latent trait of the person. The 3PL model (Birnbaum, 1968) is given as

$$P_i(\theta_j) = P(X_{ij} = 1 | \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]}, \quad (1)$$

where X_{ij} is a scored response of examinee j to item i (1 if correct and 0 if not correct), θ_j denotes the latent trait level of person j , and a_i , b_i , and c_i are the three item parameters denoting the discriminating power, the difficulty and the pseudo-guessing probability of item i , respectively. When the pseudo-guessing probability c_i is fixed to zero, the 3PL model reduces to the 2PL model. If a_i is further restricted to be equal to a constant across n items, the 3PL model reduces to the 1PL model or the Rasch model when the constant is 1.

The probability mass function of a Bernoulli random variable is expressed as

$$f(x_{ij}; P_i(\theta_j)) = P_i(\theta_j)^{x_{ij}} [1 - P_i(\theta_j)]^{1-x_{ij}}, \quad (2)$$

where x_{ij} is a realization of the random variable, X_{ij} , and $P_i(\theta_j)$ is provided in Equation 1. For item i and person j , l_0 is the natural log of the probability mass function (i.e., the log-likelihood function) and given as

$$l_{0ij} = x_{ij} \ln P_i(\theta_j) + (1 - x_{ij}) \ln [1 - P_i(\theta_j)], \quad (3)$$

where \ln denotes the natural log. l_0 is defined as

$$l_{0j} = \sum_{i=1}^n \left\{ x_{ij} \ln P_i(\theta_j) + (1 - x_{ij}) \ln [1 - P_i(\theta_j)] \right\}. \quad (4)$$

Levine and Rubin (1979) suggested that l_{0j} represents a measure of how likely it is to observe person j 's response pattern under a specific IRT model. Large negative values of l_{0j} indicate that the response pattern from person j deviates from the one predicted under the model. One issue with using l_{0j} is that the value of l_{0j} depends on θ_j ; standardizing l_{0j} removes this dependency (Drasgow et al., 1985). The standardized value, l_{zj} , is written as

$$l_{zj} = \frac{l_{0j} - E(l_{0j})}{\sqrt{Var(l_{0j})}}, \quad (5)$$

where

$$E(l_{0j}) = \sum_{i=1}^n \left\{ P_i(\theta_j) \ln P_i(\theta_j) + [1 - P_i(\theta_j)] \ln [1 - P_i(\theta_j)] \right\} \quad (6)$$

and

$$Var(l_{0j}) = \sum_{i=1}^n P_i(\theta_j) [1 - P_i(\theta_j)] \left\{ \ln \left[\frac{P_i(\theta_j)}{1 - P_i(\theta_j)} \right] \right\}^2. \quad (7)$$

In a practice setting where the item parameters in Equation 1 are assumed to be known and the latent trait parameter, θ_j , is estimated using the item parameters, the estimated trait and item parameter values are used in the Equations 4, 6, and 7 for calculating l_{zj} . As with the interpretation of l_{0j} , a large negative value of l_{zj} indicates an unusual response pattern under the chosen IRT model (i.e., misfit) while a large positive value of l_{zj} indicates overfit.

Although it was assumed that l_z follows a standard normal distribution under the null condition (i.e., item response patterns are congruent with the specified response model), several studies have found that this assumption does not hold when an estimated trait value was used (Molenaar & Hoijtink, 1990; Nering, 1995; Reise, 1995). To overcome this problem, Snijders (2001) proposed a variant of the l_z index to obtain the asymptotic standard normal distribution with estimated latent trait values instead of the true (unknown) values. This variant of l_z is referred to as the l_z^* index. When the maximum likelihood (ML) estimator is used for estimating trait values, the l_z^* expressed below is identical to Equation 5, except for the adjustment, $c_n(\hat{\theta}_j)r_i(\hat{\theta}_j)$, in the denominator:

$$l_{zj}^* = \frac{l_{0j}(\hat{\theta}_j) - E[l_{0j}(\hat{\theta}_j)]}{\sqrt{\sum_{i=1}^n \left\{ \ln \left[\frac{P_i(\hat{\theta}_j)}{1 - P_i(\hat{\theta}_j)} \right] - c_n(\hat{\theta}_j)r_i(\hat{\theta}_j) \right\}^2 P_i(\hat{\theta}_j)[1 - P_i(\hat{\theta}_j)]}}, \quad (8)$$

where $c_n(\hat{\theta}_j)$ and $r_i(\hat{\theta}_j)$ are written as

$$c_n(\hat{\theta}_j) = \frac{\sum_{i=1}^n P'_i(\hat{\theta}_j) \ln \left[\frac{P_i(\hat{\theta}_j)}{1 - P_i(\hat{\theta}_j)} \right]}{\sum_{i=1}^n P'_i(\hat{\theta}_j) r_i(\hat{\theta}_j)} \quad \text{and} \quad r_i(\hat{\theta}_j) = \frac{a_i \exp[a_i(\hat{\theta}_j - b_i)]}{c_i + \exp[a_i(\hat{\theta}_j - b_i)]}, \quad (9)$$

where $P'_i(\hat{\theta}_j)$ is the first derivative of $P_i(\theta)$ with respect to θ evaluated at the trait estimate for person j . When the Rasch model or the 2PL model is used, $r_i(\hat{\theta}_j)$ will be greatly simplified. The interpretation of l_z^* is identical to that of l_z : larger negative values indicate misfit and large positive values indicate overfit.

CUSUM

CUSUM has been widely adapted in modern industries as one of statistical process control methods to monitor change detection. CUSUM is designed to detect deviations from acceptable criteria (see, e.g., Montgomery, 2012). Bradlow, Weiss, and Cho (1998) and van Krimpen-Stoop and Meijer (2001) applied CUSUM to person-fit research. A residual value, $X_{ij} - P_i(\theta_j)$, is calculated for a scored response (X_{ij}) of item i from person j , where $P_i(\theta_j)$ is the model predicted probability. This residual is accumulated over the set of items person j received. CUSUM is defined as

$$C_{ij}^+ = \max(0, C_{i-1j}^+ + T_{ij}) \quad (10)$$

and

$$C_{ij}^- = \min(0, C_{i-1j}^- + T_{ij}) \quad (11)$$

with

$$T_{ij} = \frac{[X_{ij} - P_i(\theta_j)]}{n_j}; C_{0j}^+ = 0; C_{0j}^- = 0, \quad (12)$$

where T_{ij} is the residual value corrected for the test length and n_j is the number of items person j took. For a fixed form test where the items do not vary across persons, n_j is a constant. However, n_j can be different values across persons in computerized adaptive testing (CAT).

C^+ and C^- accumulate positive and negative residuals, respectively. C^+ becomes large only for a set of consecutive positive residuals, whereas C^- becomes small only for a set of consecutive negative residuals. Two boundary values are required to classify a response pattern into aberrant and nonaberrant categories. Let UB be an upper boundary value and LB be the lower boundary value. If $C^+ > UB$ or $C^- < LB$ then the item response pattern is classified as misfitting; if not, the pattern is classified as fitting. The boundary values can be obtained using simulated data (e.g., Meijer, 2002; van Krimpen-Stoop & Meijer, 2002).

Armstrong and Shi (2009) argued that one drawback of most of existing person-fit indexes is that they do not utilize item sequencing information in their computation. Ignoring item sequencing information leads to the inability to detect aberrant patterns because when looking at response patterns only in the aggregate, unusual strings of positive (or negative) residuals are often cancelled by sets of negative (or positive) residuals elsewhere in the exam. However, CUSUM can detect aberrant patterns over subsets of items because CUSUM reports information on aberrant patterns as they occur.

Nonparametric Person-Fit Indexes

A nonparametric person-fit index does not require estimated IRT model parameters but relies on classical test theory or Guttman scales to derive expected responses or response patterns. Karabatsos (2003) examined 36 person-fit statistics in his study. H^T and $U3$ were two of the top performing indexes in identifying aberrant response patterns.

U3

Van der Flier (1982) developed the $U3$ index in the context of the nonparametric Mokken monotone homogeneity model (Mokken, 1971), where item response functions are nondecreasing, but they may cross. When each of N persons took a n -item multiple-choice test, the scored data matrix consists of 0s and 1s with N rows and n columns. The columns are reordered with the easiest item on the first column and the hardest item on the last column (i.e., n th column). The $U3$ index is defined as

$$U3_j = \frac{\sum_{i=1}^{r_j} \ln\left(\frac{\pi_i}{1-\pi_i}\right) - \sum_{i=1}^n X_{ij} \ln\left(\frac{\pi_i}{1-\pi_i}\right)}{\sum_{i=1}^{r_j} \ln\left(\frac{\pi_i}{1-\pi_i}\right) - \sum_{i=n-r_j+1}^n \ln\left(\frac{\pi_i}{1-\pi_i}\right)}, \quad (13)$$

where π_i is the proportion correct or the average value of item i of the reordered data matrix, X_{ij} is the item response for item i and person j , and r_j is the total score for person j .

The $U3$ index represents the degree to which a person's response vector deviates from the Guttman scale: 0 for the response vector being consistent with the Guttman scale and 1 for the opposite. A critical value can be obtained either using simulation or by selecting empirical cut-off values (e.g., 95% quantile).

H^T

Sijtsma (1986) proposed the person-fit index, H^T . The index is a correlation index that measures the similarity between the response vector of person j and the response vectors of the rest of persons in a group to which person j belongs. In essence, it indicates the degree to which a person's response pattern on a set of items deviates from another person's response pattern. H^T for person j is calculated as

$$H^T = \frac{\sum_{k \neq j} \beta_{jk} - \beta_j \beta_k}{\sum_{k \neq j} \beta_j (1 - \beta_k)}, \quad (14)$$

where β_j is the proportion of correct items for person j , β_{jk} is the proportion of correct items for both person j and person k , and the denominator represents the maximum possible covariance when assuming that $j < k$ implies $\beta_j \leq \beta_k$ (see Sijtsma & Meijer, 1992).

The range of H^T is -1 to $+1$. A positive value of H^T indicates a response vector that is consistent with other examinees' response patterns and a negative value is indicative

of an aberrant response pattern. Karabatsos (2003) suggested that the critical value $H^T = .22$ best identified aberrant respondents. As with $U3$, a critical value can be obtained using either simulation or empirically, as long as there are not many aberrant respondents.

Data Analysis Using Person-Fit Indexes

It is important to check whether the response data are influenced by only one underlying latent dimension before calculating any person-fit index. Any person misfit could be caused by the violation of the unidimensionality assumption rather than by aberrant response patterns. Dimensionality assessment was conducted using *sirt* R package (Robitzsch, 2015), which implements the algorithm of the DETECT program (Stout et al., 1996; Zhang & Stout, 1999a, 1999b). According to Zhang (2007), a D value that is less than .2 indicates essential unidimensionality. The D values of Form 1 and Form 2 are -0.089 and -0.090 , respectively. It appears that the unidimensionality assumption is not violated in either form.

All parametric person fit indexes rely on IRT. To reap the benefits of IRT (e.g., persons and items on the same scale, person abilities are independent of item difficulties, and score precision) one must verify a model-data fit. One facet of the model-data fit investigation involves assessing the assumption of local independence. Fisher's r -to- z transformed Q_3 was used to assess the local independence assumption. Most of the transformed Q_3 values fall between the critical values (-2.56 and 2.56) in Figure 4.2. Although there are quite a few transformed Q_3 values that fall outside the critical values, it appears that the local independence assumption is not violated in either form considering the fact that the Type I error rates for the transformed Q_3 was quite liberal in some studies (see Kim, De Ayala, Ferdous, & Nering, 2011).

Four person-fit indexes (l_z , l_z^* , $U3$, and H^T) were calculated using *PerFit* R package (Tendeiro, 2014). R code was created to calculate CUSUM values. Item and person parameter estimates are required for the three parametric person-fit indexes (l_z , l_z^* , and CUSUM). The common data sets provide Rasch difficulty values (i.e., item bank values). However, both item and person parameters were recalibrated using the responses in

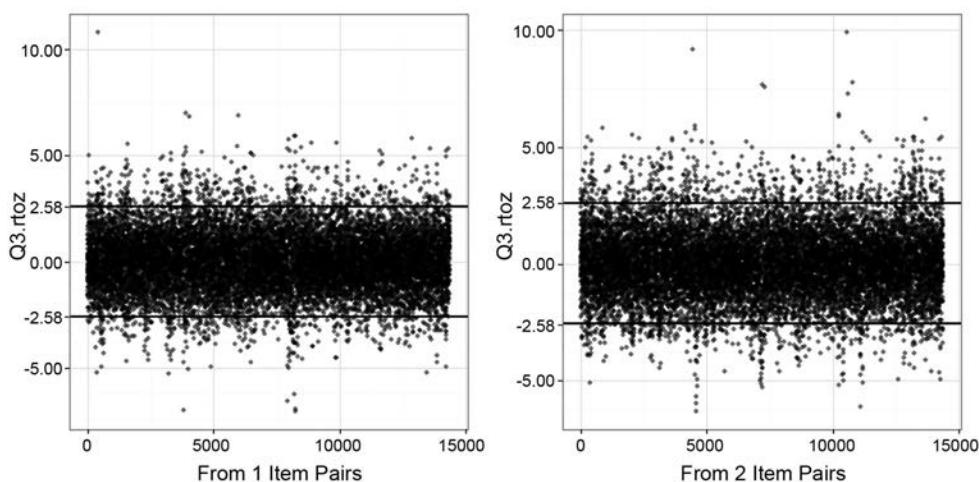


Figure 4.2 Scatter Plots for Fisher's r -to- z Transformed Q_3

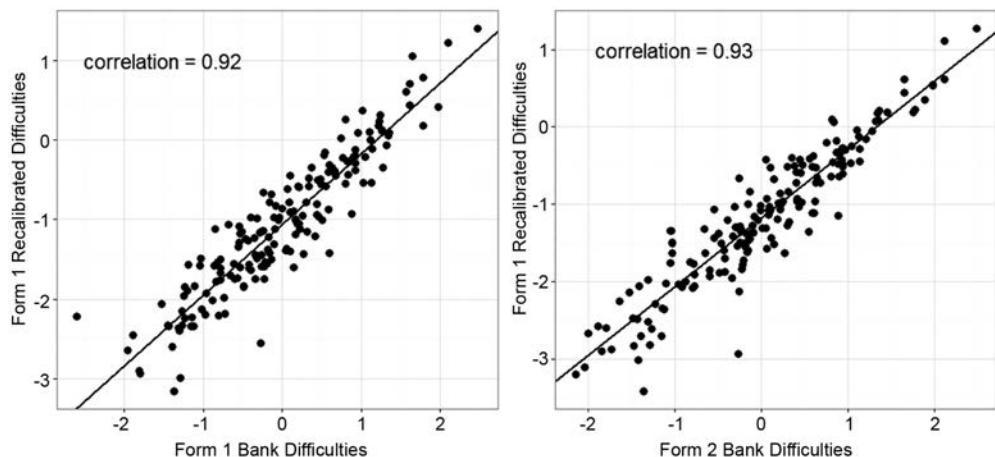


Figure 4.3 Comparison of Item Bank Difficulties and Recalibrated Difficulties

the common data sets because the performances of the three parametric person-fit indexes could be impacted by the accuracy of the estimates. Figure 4.3 shows the bank item and recalibrated item difficulties and their correlation for both forms.

Both l_z and l_z^* asymptotic distributions are assumed to follow the standard normal distribution. Figure 4.4 displays how close the empirical density distributions of these two indexes are close to the standard normal distribution. The l_z distributions are leptokurtic in both forms, whereas the l_z^* distributions are mesokurtic (that is, similar to a normal distribution). It appears that both l_z and l_z^* distributions have heavy tails on the left side, which could be a problem because these two indexes classify test takers with small index values as misfitting when using cutoff values such as -1.96 or -2.58 .

Figure 4.5 shows the density distributions for $U3$ and H^T with their theoretical normal distributions superimposed. The mean and standard deviation of each superimposed normal distribution were made to have same values as its empirical distribution. It appears that each empirical density distribution follows a normal distribution. As both l_z and l_z^* distributions do not closely follow the standard normal distribution and there is no suggested cutoff value for $U3$ and H^T , depending on which index was used, either upper or lower 5% quantile value was used for flagging test takers (see Figure 4.6). To reduce false positive rate, the flagging criterion can be adjusted (e.g., upper or lower 1% quantile).

There is no suggested cutoff value for CUSUM. Researchers have recommended conducting simulation studies under a null condition (i.e., generating responses consistent with IRT models) to derive empirical cutoff values (see Meijer, 2002; van Krimpen-Stoop & Meijer, 2002). Here, the recalibrated item parameter estimates and subsequently estimated ability values were used for generating item responses under the null condition with 100 replications for each form. In Figure 4.7, the left panels show one of the 100 replications with two cutoff lines overlaid (0.051 for CUSUM C^+ and -0.051 for CUSUM C^-), and the middle and right panels show Form 1 and Form 2 CUSUM charts, respectively. A test taker whose CUSUM values were either larger than 0.051 or smaller than -0.051 were flagged.

The common data sets include flagged test takers who engaged in fraudulent test behaviors and were identified by the data provider. It is known that at least some of these test takers illegally obtained test items prior to the exam (i.e., item preknowledge) though other types of misconduct were possible as well. There are 46 flagged test takers

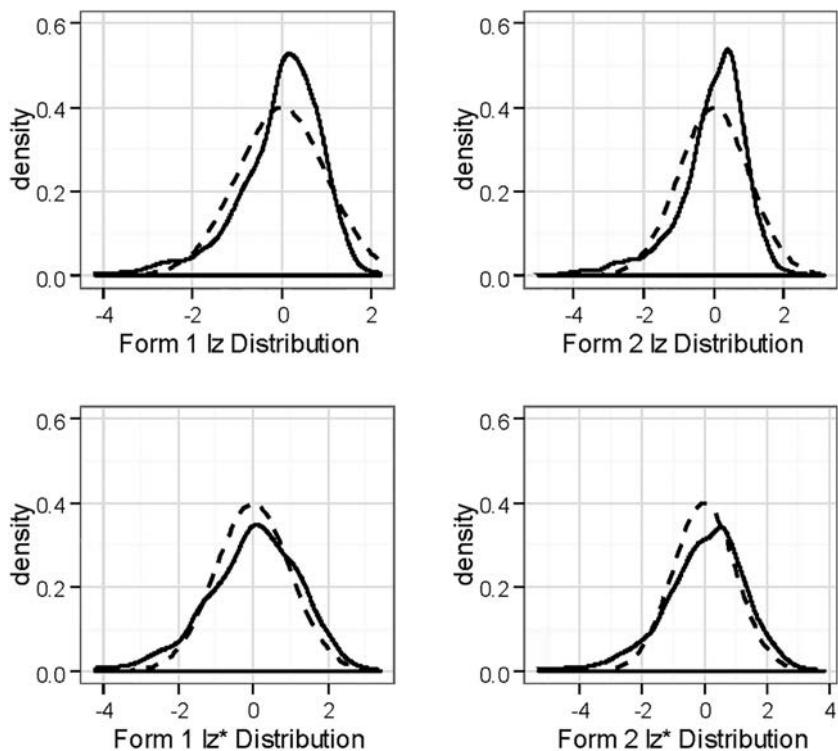


Figure 4.4 Comparison of I_z and I_z^* Distributions (solid lines) and the Standard Normal Distribution (dashed line)

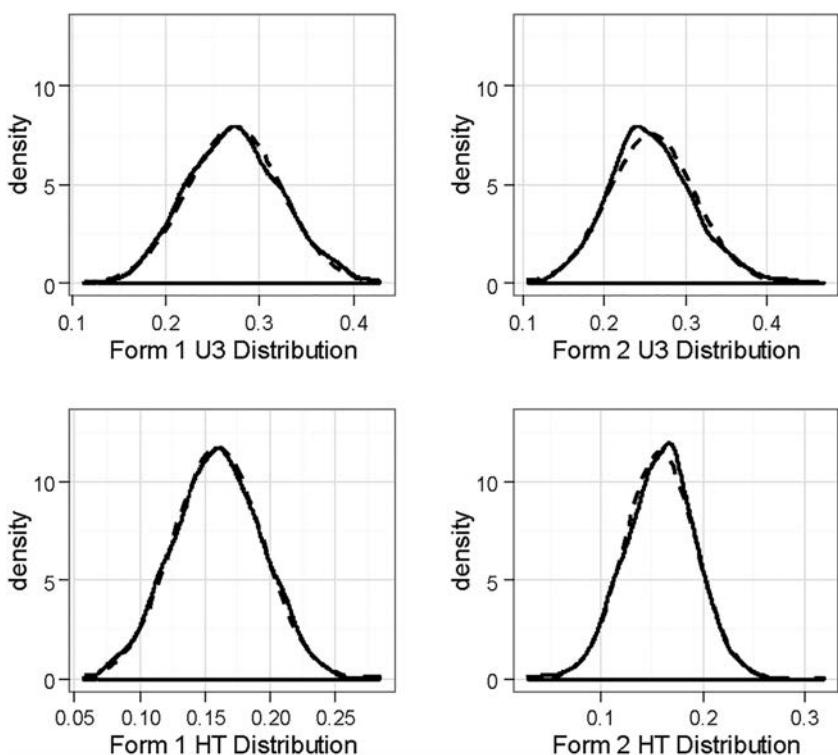


Figure 4.5 Comparison of U_3 and H^T Distributions (solid lines) and Their Theoretical Normal Distribution (dashed line)

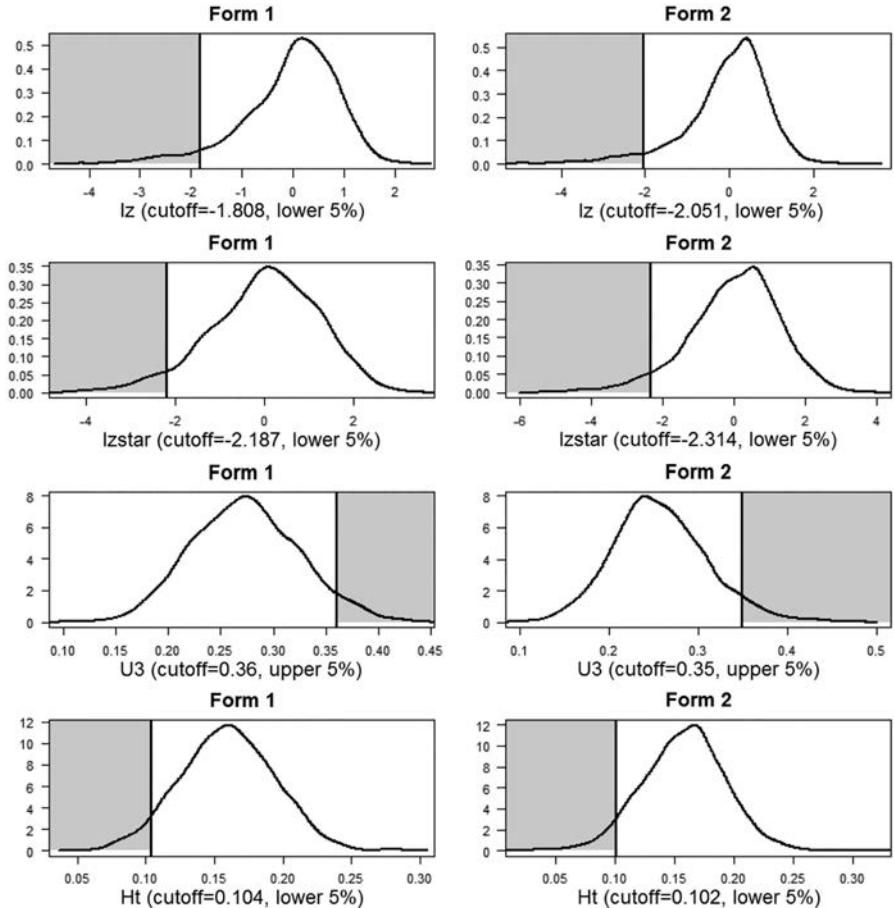


Figure 4.6 Cutoff Values for I_z , I_z^* , U_3 and H^T

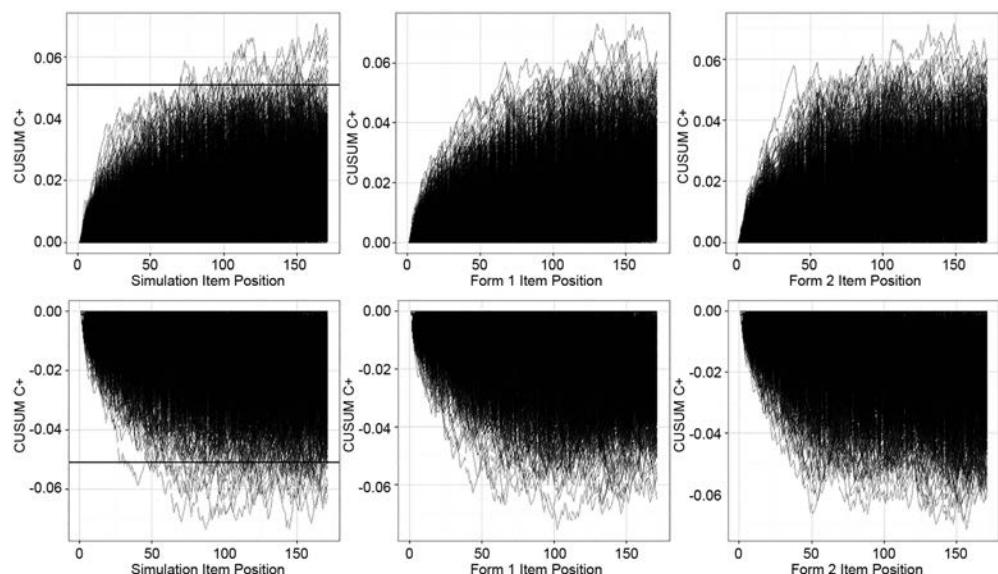


Figure 4.7 Simulated CUSUM and Empirical CUSUM charts

Note: The thick horizontal lines on the simulated CUSUM charts indicate empirical cutoff values: -0.051 for CUSUM C^+ and 0.051 for CUSUM C^- .

on Form 1 and 48 on Form 2. These numbers of the flagged test takers represent about 2.8% of all Form 1 test takers and 2.9% of all Form 2 test takers. As either upper or lower 5% quantile values were used for flagging test takers using the six person-fit indexes, the flagged test takers by each of the six indexes will be about 5% of the test takers for each form. Table 4.1 shows the number of flagged cases for each index. As expected, there was at least 4.95% of the test takers flagged by one of the six indexes. It appears that CUSUM C^+ flagged slightly more test takers than did the other indexes. Some of test takers flagged by CUSUM C^- appear to be very slow starters (e.g., very severe warming-up effect). Their performances became better after 100 items (see Figure 4.8).

It is interesting to see how many of the flagged test takers were also identified as aberrant test takers by the six person-fit indexes. Table 4.2 and Table 4.3 provide a snapshot of the result.

Table 4.1 Numbers of Test Takers Identified by Six Person-Fit Indexes

Forms	Flagged	I_z	I_z^*	$U3$	H^T	CUSUM C^+	CUSUM C^-
Form 1	46	82	82	83	81	88	106
Form 2	48	83	83	85	84	93	144

Note: Flagged—the common data sets include flagged test takers that were identified by the data provider.

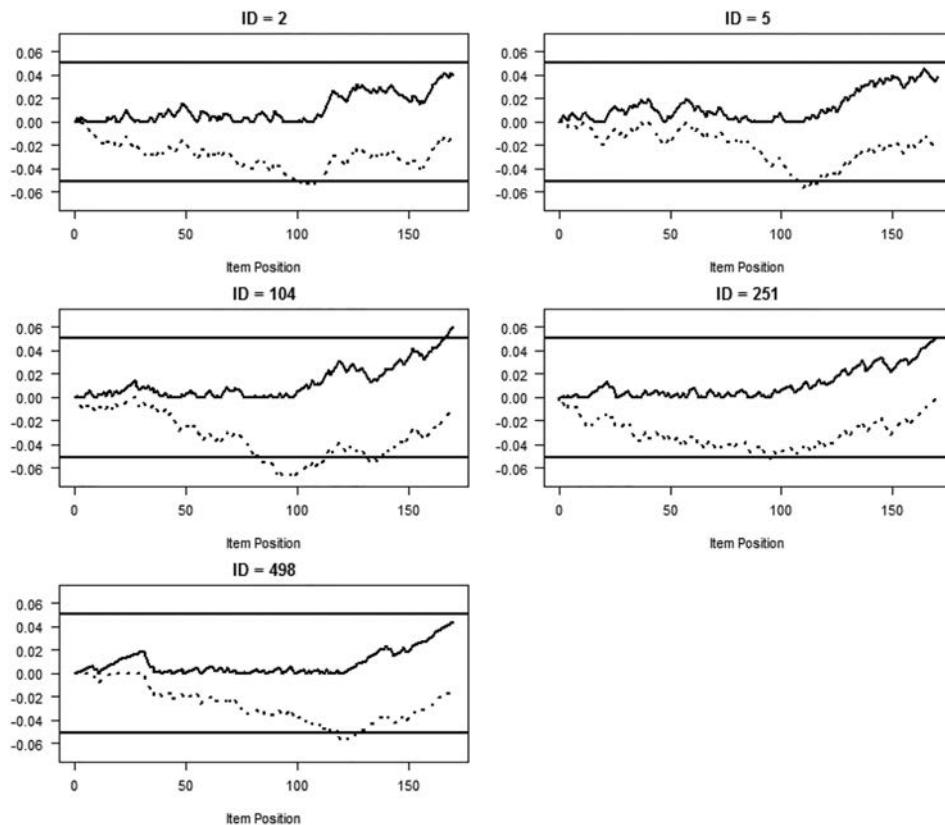


Figure 4.8 CUSUM Charts for Flagged Test Takers by Both the Data Provider and CUSUM C^-

Note: Solid lines represent CUSUM C^+ and the dotted lines CUSUM C^- .

Table 4.2 Form 1 Snapshot Comparison of Flagged Cases by the Data Provider and Person-Fit Indexes

EID	Flagged	l_z	l_z^*	U3	H^T	CUSUM C^+	CUSUM C^-
e100002	X	X	X	X	X		X
e100003	X	X	X	X			
e100004	X	X	X	X	X		
e100005	X	X	X	X			X
e100011	X	X	X				
e100103	X						
e100104	X				X		X
e100141	X						
e100149	X						
e100191	X						
e100219	X						
e100226	X						
e100248	X						
e100251	X			X	X	X	X
e100271	X						
e100292	X						
e100415	X						
e100452	X						
e100453	X						
e100478	X						
e100494	X						
e100498	X						X
e100505	X						
e100524	X						
e100623	X						
e100624	X						
e100639	X						
e100645	X						
e100687	X						
e100892	X						
e100903	X						
e100919	X						
e100921	X						
e100923	X						
e100983	X						
e101176	X						
e101267	X						
e101307	X						
e101356	X						
e101358	X						
e101413	X						
e101426	X						
e101432	X						
e101450	X						
e101585	X						
e101620	X						

Table 4.3 Form 2 Snapshot Comparison of Flagged Cases by the Data Provider and Person-Fit Indexes

EID	Flagged	l_z	l_z^*	U3	H^T	CUSUM C ⁺	CUSUM C ⁻
e200001	X	X	X	X	X		
e200002	X	X	X	X	X		X
e200007	X	X	X	X			
e200008	X	X	X	X	X		
e200025	X	X	X	X			
e200336	X						X
e200352	X						X
e200368	X						
e200407	X						X
e200417	X		X	X	X		
e200438	X						
e200448	X						
e200503	X						
e200512	X						
e200528	X				X		
e200530	X						
e200545	X						
e200551	X						
e200584	X						
e200592	X						
e200611	X						
e200663	X						
e200682	X						
e200686	X						
e200695	X				X		
e200702	X						
e200703	X						
e200709	X						
e200837	X						
e200873	X						
e200894	X						
e200900	X						
e200904	X						
e200911	X						
e200949	X						
e200981	X				X		
e201011	X						
e201039	X		X	X	X	X	
e201058	X						
e201094	X						
e201158	X						
e201173	X						
e201422	X				X		
e201481	X						
e201487	X						
e201504	X						
e201616	X						
e201627	X						

It appears that the flagging patterns of I_z , I_z^* , and $U3$ are similar one another: one test taker flagged by one of these indexes is more likely to be flagged by the other indexes. Although there were test takers flagged by both CUSUM C^+ and CUSUM C^- , these two CUSUM statistics tend to flag different test takers. Figure 4.9 shows why these flagging

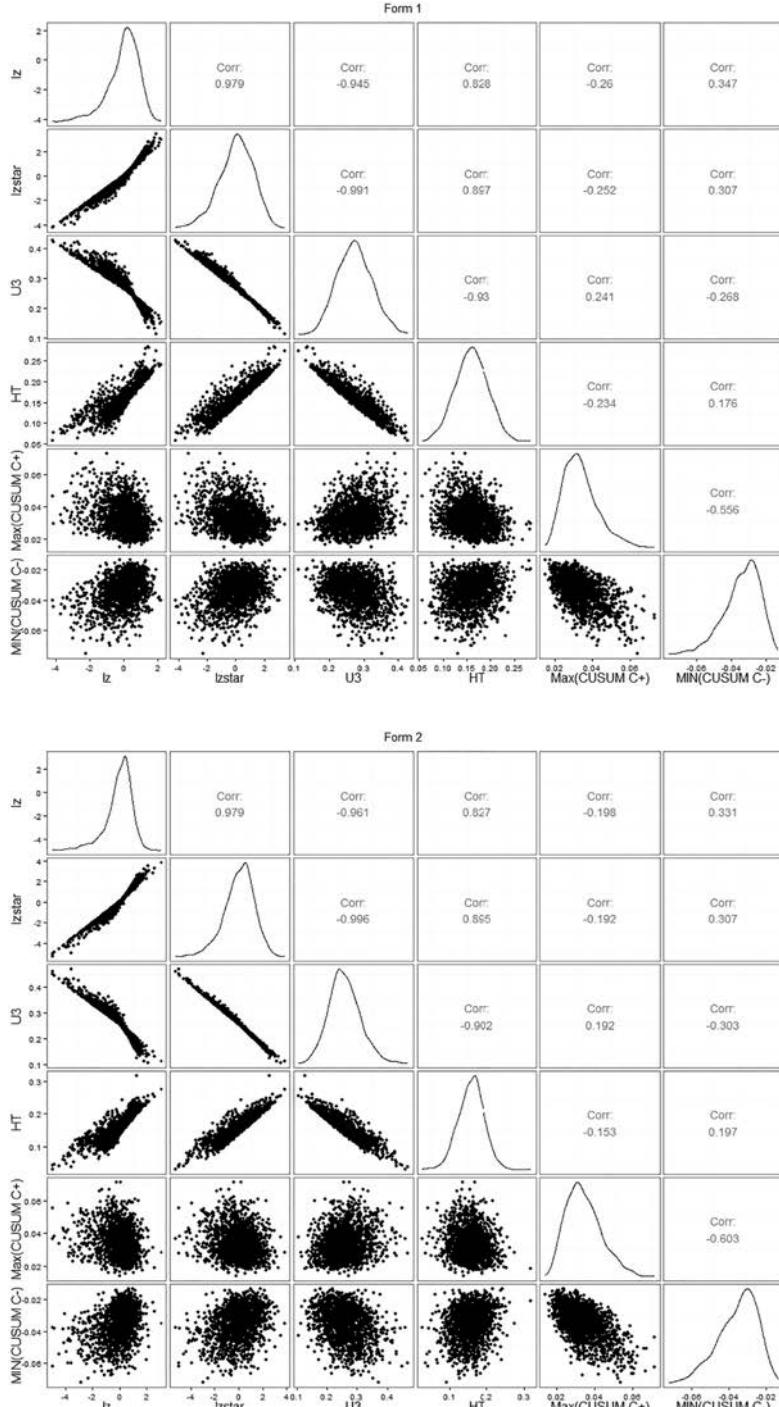


Figure 4.9 Scatter Plots and Correlations Among the Person-Fit Statistics: I_z , I_z^* , $U3$, HT , $\text{Max}(\text{CUSUM } C^+)$, and $\text{Min}(\text{CUSUM } C^-)$

patterns occurred. The correlations among l_z , l_z^* , and $U3$ were high. The correlations between the two CUSUM indexes and were relatively low. The maximum value of CUSUM C^+ and the minimum value of CUSUM C^- for each test taker were used in Figure 4.9 because there are as many CUSUM C^+ and CUSUM C^- values as there are the items on each form.

In sum, the flagged test takers by l_z and l_z^* are almost identical. Thus, it does not matter much which of the two indexes practitioners use considering the fact that the empirical density distribution of l_z^* does not follow the standardized normal distribution (e.g., the heavy left tail). $U3$ is promising in that it does not require IRT parameter estimates and its performance is on par with the two parametric indexes. CUSUM charts could be helpful in providing insights into the test taking processes. As would be expected, all the six person-fit indexes failed to identify all the flagged cases in the common data set. As stated previously, there are many factors causing aberrant response patterns. It is unknown how the data provider flagged the test takers; however, it is believed that some of these test takers had item preknowledge. In the following section, two response time models are explained and explored as to how response time models help identify aberrant response time patterns.

Response Time Models

In this chapter, we utilize two RT models to detect item preknowledge. One is the effective response time model (Meijer & Sotaridona, 2006) and the other is the hierarchical framework for response times and item responses proposed by van der Linden (2007). The former was developed for specifically addressing the issue of item preknowledge. The latter was initially proposed for increasing the accuracy of item parameter estimation when using response times simultaneously. However, residual RTs calculated from the hierarchical framework can be used for detecting item preknowledge or aberrant RT patterns.

Effective Response Time Model

Meijer and Sotaridona (2006) proposed a response time model to detect preknowledge of items on CBT. The authors defined the “effective response time” (ERT) as the time spent by person j answering item i correctly. Both item response and RT models are required to estimate ERT for each item. The item response model for dichotomously scored items is specified in Equation 1. Meijer and Sotaridona (2006) used a loglinear model discussed in van der Linden and van Krimpen-Stoop (2003) to estimate a parameter for the slowness of person j . The loglinear model specifies RTs on the log scale as

$$\ln T_{ij} = \mu + \delta_i + \tau_j + \varepsilon_{ij} \quad (15)$$

with $\varepsilon_{ij} \sim N(0, \sigma^2)$,

where μ is a parameter for the mean of log response time for n items and N persons, δ_i is a parameter for the response time required for item i , τ_j is a parameter indicating the slowness of person j , and ε_{ij} is a residual assumed to follow a normal distribution with a mean of 0 and a standard deviation σ . The slowness parameter of person j is calculated as

$$\tau_j \equiv E_i(\ln T_{ij}) - E_{\bar{i}}(\ln T_{ij}) \quad (16)$$

where $E_i(\ln T_{ij})$ is the mean of person j 's log response time over n items, and $E_{ij}(\ln T_{ij})$ is equal to μ . The ERT model is similar to the loglinear model in Equation 15 but with a different parameterization. The ERT is specified as

$$\ln T_{ij} = \beta_0 + \beta_1 \theta_j + \beta_2 \tau_j + \varepsilon_j \quad (17)$$

$$\text{with } \varepsilon_j \sim N(0, \sigma_i^2),$$

where β_0 , β_1 , and β_2 are regression coefficients, ε_j is an error term, and both θ_j and τ_j are two known regressors coming from the ability estimation process using known item parameters in Equation 1 and from Equation 16, respectively.

There are two unique characteristics of calculating ERT. First, the ERT is estimated for each item to remove the effect of item characteristics on estimating ERT. Second, when estimating the regression coefficients using Equation 17, the RT data from selected persons meeting two criteria are used: (1) a person correctly responding to item i and (2) a person whose probability of a correct response on item i is larger than a prespecified value (e.g., 0.25). These characteristics result in a different set of persons for estimating regression coefficients for each item. The selection criteria reduce the variance in the observed response time by removing examinees such as guessers who didn't spend much time considering the item, which helps establish a valid ERT for each item.

Meijer and Sotaridona (2006) suggested that the difference between observed response times and predicted response times (i.e., ERT) from Equation 17 can be used for identifying item preknowledge. As it is assumed that the error term is normally distributed with a mean of 0 and a standard deviation of σ_i , an item-level z score for RT aberrance can be calculated for examinee p suspected of having preknowledge for item i as

$$z_{ip} = \frac{\ln T_{ip} - \widehat{\ln T_{ij}}}{\sigma_i}, \quad (18)$$

where $\sigma_i^2 = (N_i - 1)^{-1} \sum_j^{N_i} (\ln T_{ip} - \widehat{\ln T_{ij}})^2$ is the variance of the log response time for item i and N_i is the number of persons selected for item i using the two criteria. To compute a person-level statistic, the sum of the squares of the z statistics for n items is calculated. The person-level statistic follows a chi-squared distribution with n degrees of freedom. A person with a large chi-squared value can be flagged. Further investigation of the flagged persons using the ERT model must be combined with other evidence, such as the deviation between the observed and expected response.

Hierarchical Framework for Response Times and Item Responses

Van der Linden (2007) introduced a general hierarchical framework for simultaneously modeling both item responses and response time. An item response model (e.g., the model in Equation (1)) and a response time model are specified at the first level of the hierarchical framework. A lognormal model has been used frequently for response times (Thissen, 1983; van der Linden, Scrams, & Schnipke, 1999). Van der Linden (2006, 2007) discussed that the distribution of the log of response time follows

a normal distribution with a mean of $\beta_i - \tau_j$ and a standard deviation of $1/\alpha_i$. The model can be described as

$$T_{ij} \sim f(t_{ij}; \tau_j, \alpha_i, \beta_i) \quad (19)$$

$$\text{with } f(t_{ij}; \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_i(\ln t_{ij} - (\beta_i - \tau_j))\right]^2\right\},$$

where t_{ij} , the realization of a random variable T_{ij} , is the response time spent on item i by person j and $\ln t_{ij}$ is the logarithm of t_{ij} , τ_j is the person parameter for the speed of person j , and α_i and β_i are the item parameters of item i representing the discriminating power and the time intensity of item i , respectively. The smaller τ is associated with larger amounts of time spent on items, while the smaller β , the more quickly persons responded to an item. Note that α is inversely related to the standard deviation of the log normal distribution. That is, smaller values of α are associated with more dispersed log normal response time distributions of an item, which means that the item is less discriminating among persons with different speed parameters. When α_i is fixed to one across n items, the mean and standard deviation of the lognormal response time distribution are $\beta_i - \tau_j$ and 1, respectively.

A bivariate normal distribution for the person parameters from the first-level models is specified at the second-level. The model can be written as

$$f(\theta, \tau; \mu_\theta, \mu_\tau, \sigma_\theta, \sigma_\tau, \rho_{\theta\tau}) = \frac{1}{2\pi\sigma_\theta\sigma_\tau\sqrt{1-\rho_{\theta\tau}^2}} \exp\left[-\frac{1}{2(1-\rho_{\theta\tau}^2)}(z_\theta^2 - 2\rho_{\theta\tau}z_\theta z_\tau + z_\tau^2)\right] \quad (20)$$

$$\text{with } z_\theta = (\theta - \mu_\theta)/\sigma_\theta \text{ and } z_\tau = (\tau - \mu_\tau)/\sigma_\tau,$$

where μ_θ and μ_τ are the means of the ability parameters and speed parameters, σ_θ and σ_τ are the standard deviations of these parameters, and $\rho_{\theta\tau}$ is their correlation. A multivariate normal distribution for the item parameters from the first-level models is specified at the second level in the following form:

$$f(\xi_i; \mu, \Sigma) = \frac{|\Sigma^{-1}|^{1/2}}{(2\pi)^{5/2}} \exp\left[-\frac{1}{2}(\xi_i - \mu)^T \Sigma^{-1} (\xi_i - \mu)\right], \quad (21)$$

where ξ_i is the vector of the item parameters ($a_i, b_i, c_i, \alpha_i, \beta_i$), μ is the mean of the item parameters ($\mu_a, \mu_b, \mu_c, \mu_\alpha, \mu_\beta$), and Σ is the variance-covariance matrix of the item parameters.

Residual RTs calculated from the hierarchical framework can be used for detecting item preknowledge or aberrant RT patterns. Once the parameters are estimated, the fact that the distribution of the log of response time follows a normal distribution with a mean of $\beta_i - \tau_j$ and a standard deviation of $1/\alpha_i$ can be used for flagging cases. More specifically, for each item, a standard normal deviates can be calculated by dividing the difference between the log of response time and the mean by the standard deviation. The sum of squares of standard normal deviates over items follows a chi-squared distribution with the degrees of freedom that is equal to the number of the standard normal deviates being summed. A test taker whose chi-squared value is larger than a critical value is flagged.

Data Analysis Using Response Time Models

R code was created to estimate the parameters of the ERT model. z_{ip} statistics in Equation 18 were calculated using the estimated parameters. The sum of the squares of the z_{ip} statistics over n items follows a chi-squared distribution with n degrees of freedom. A person with a large chi-squared value can be flagged. The critical value of chi-squared with 170 degrees of freedom at the 5% significance level is 201.42. There were 34.0% of Form 1 test takers whose chi-squared values were larger than the critical value, and 33.4% of Form 2 test takers. Figure 4.10 shows two scatter plots of chi-squared values for the test takers for both forms. The flagged test takers by the data provider were plotted with a larger solid dot. As can be seen, some of the flagged test takers have extremely large chi-squared values (e.g., chi-squared value > 500). However, it is not easy to separate the majority of the flagged test takers from the unflagged test takers.

Two basic assumptions underlying the hierarchical framework (van der Linden, 2007) need to be validated before estimating model parameters. One is that the distribution of log response times follows a normal distribution, and the other is that a test taker operates at a constant speed. Figure 4.11 displays Form 1 empirical density distribution of log response times and a superimposed theoretical normal distribution with mean equal to the mean of the log response times and standard deviation equal to the standard deviation of the log response times. Figure 4.12 shows the log response times of 170 operational items for each test taker who received Form 1. A white solid line passes through the average value of the log response times for each item. It appears that Figure 4.11 and Figure 4.12 support the normality assumption and the speed constancy assumption, respectively.

The hierarchical model item and person parameters for each form were estimated using the Gibbs sampler implemented in *cirt* R package (Fox, Entink, & van der Linden, 2007). The item response model and response time model specified at the first level were the Rasch model and the one-parameter RT model, respectively. Figure 4.13

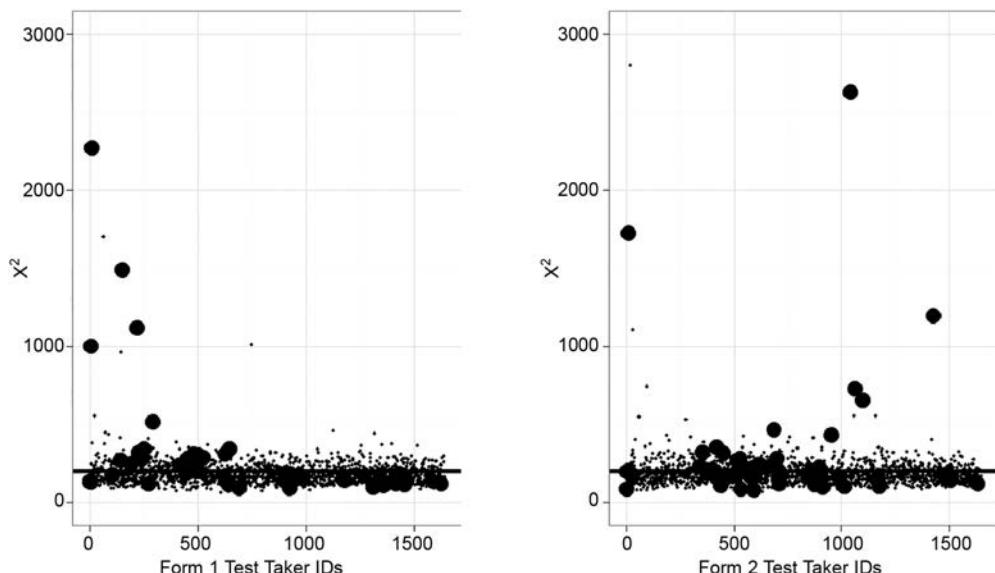


Figure 4.10 Scatter Plots for ERT χ^2 Values

Note: The thick horizontal lines indicate the critical value (201.42).

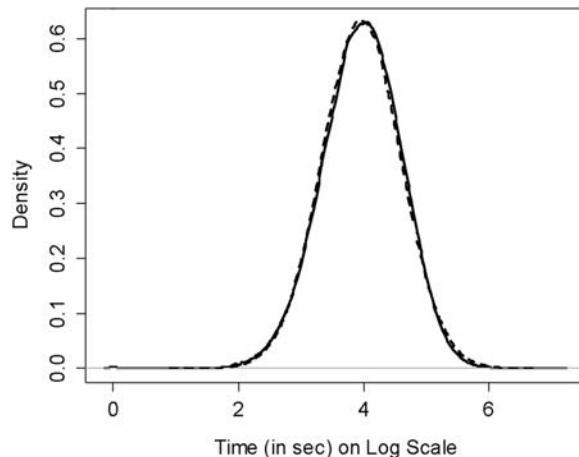


Figure 4.11 Form 1 Distribution of Log Response Times (solid line) and an Overlaying Normal Distribution (dashed line)

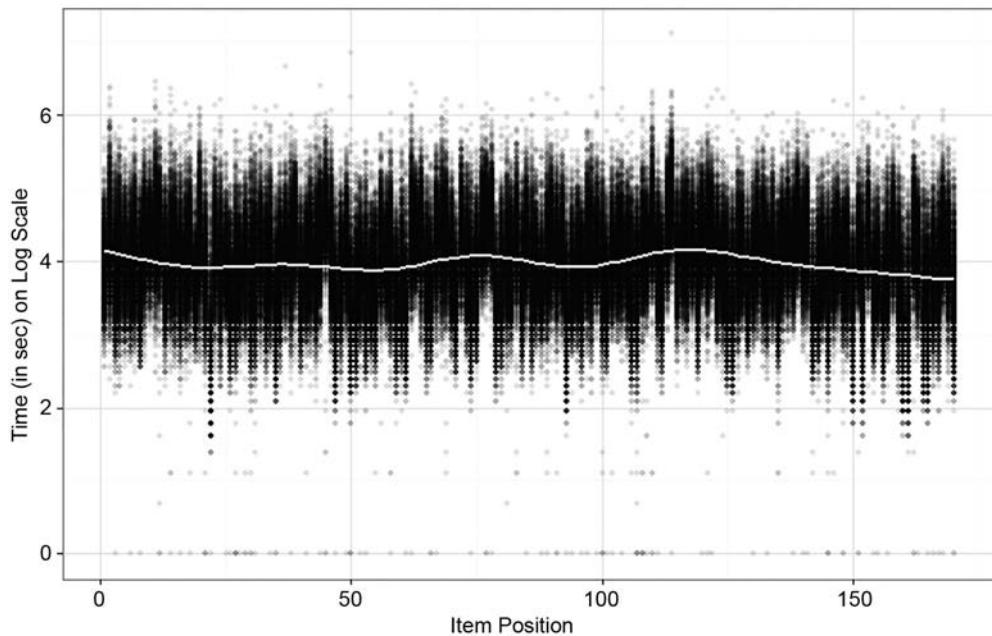


Figure 4.12 Form 1 Log Response Times with Average Speed of 4 Seconds (on the natural log scale) or About 60 Seconds per an Item

shows that both item intensity parameter estimates (i.e., difficulty value) and person speed parameter estimates from the model correspond well with the values calculated from the data. A test taker with a large speed estimate tended to spend less time on each item than a test taker with a small speed estimate. A test taker tended to spend more time on an item with a large item time intensity estimate than on an item with a small item intensity estimate.

After verifying a strong model-data fit, the predicted time, $\beta_i - \tau_j$, in Equation 19 was calculated for item i and person j . Subsequently, the residual RTs can be

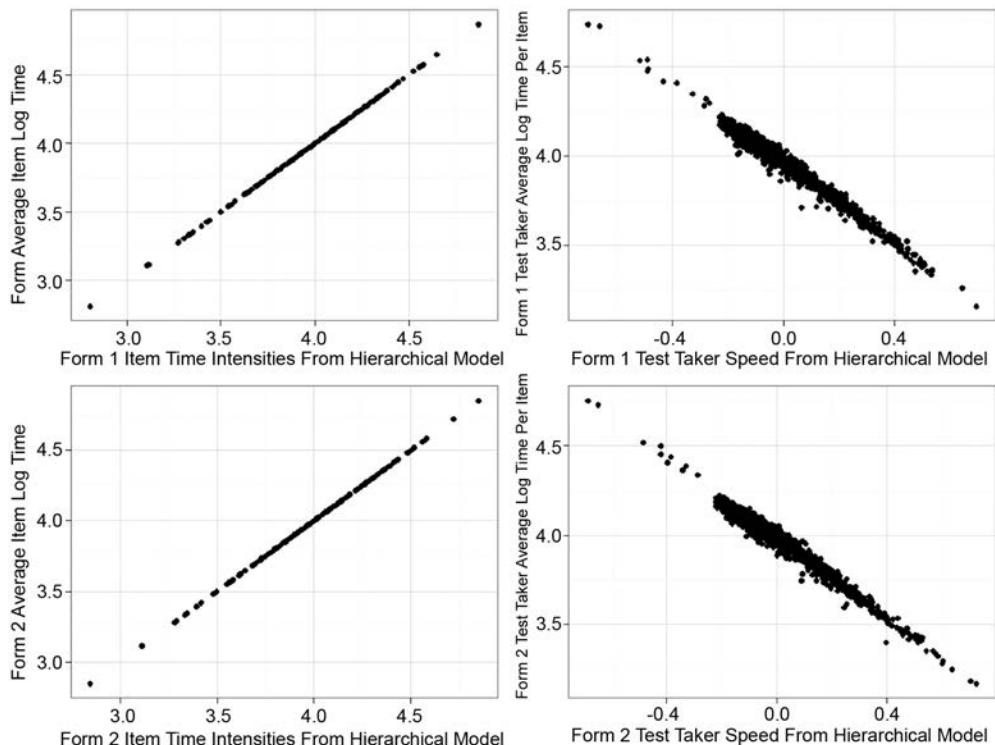


Figure 4.13 Scatter Plots for RT-Related Item and Person Parameter Estimates with RT Data

calculated: $\ln t_{ij} - (\beta_i - \tau_j)$. As with the ERT model, the sum of the squares of the residual RTs over n items follows a chi-squared distribution with n degrees of freedom because the distribution of the log of response time follows a normal distribution with a mean of $(\beta_i - \tau_j)$ and a standard deviation of 1. Figure 4.14 shows two scatter plots of chi-squared values for the test takers for both forms. The patterns in Figure 4.14 are nearly identical the ones in Figure 4.10. The flagged test takers by the data provider were again plotted with a larger solid dot. As with ERT, it is not easy to separate the majority of the flagged test takers from the unflagged test takers.

In sum, both models provided similar patterns of chi-squared values. When the critical value was applied to the chi-squared values from the ERT model, the flagging criterion was too liberal. It flagged about one-third of the test takers. When the same critical value was applied to the chi-squared values from the hierarchical framework model, the flagging criterion appeared too conservative: seven and five test takers were flagged from Form 1 and Form 2, respectively. There were four flagged test takers by the data provider among the seven from Form 1, and three among the five from Form 2.

It appears that using chi-squared tests was not effective for identifying aberrant response time patterns, considering that the ERT model was too liberal and the hierarchical model was too conservative. The process of summing all squares of the residuals may make it difficult to identify aberrant response patterns observed in a small set of items. In addition, the sign of a residual is, in and of itself, important, but squaring will remove any negative sign. The sign of a residual indicates the type of deviation: a positive sign if observed response time exceeded expected (or model-predicted)

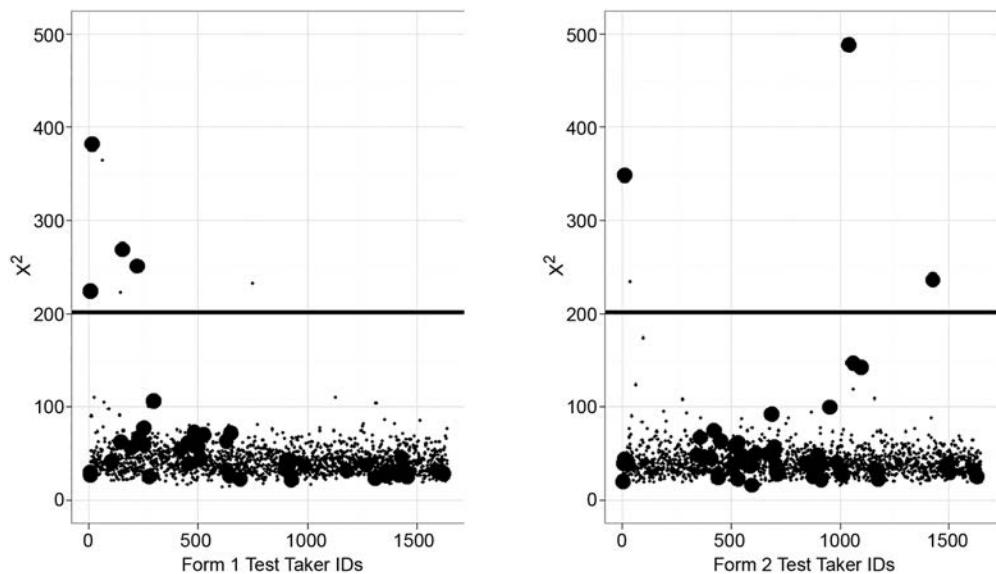


Figure 4.14 Scatter Plots for the Hierarchical Framework χ^2 Values. The Thick Horizontal Lines Indicate the Critical Value (201.42)

response time and the opposite applies to a negative sign. The two types of deviations may represent different testing aberrance (e.g., a negative deviation was observed with item preknowledge, a positive with item harvesting). It could be interesting to apply a CUSUM-type calculation to the residuals since CUSUM charts can detect aberrant response time patterns at the item level.

MACHINE LEARNING-BASED APPROACH

Machine learning can be defined as computational methods using past information for improving performance or having accurate predictions (Mohri, Rostamizadeh, & Talwalkar, 2012). Machine learning algorithms are applied to diverse areas such as classification, clustering, regression, and dimensionality reduction. The recent development in the machine learning field coupled with the continuing growth in the amount of computing power provides a vast set of tools for understanding data. Machine learning tools can be categorized as either *supervised* or *unsupervised* (James, Witten, Hastie, & Tibshirani, 2013).

Most machine learning algorithms are supervised. The main goal of the supervised algorithms is to construct predictive models for classification and prediction tasks. In the supervised learning, a training data set consists of a set of p feature variables $X_1, X_2, X_3, \dots, X_p$ paired with a target variable Y for each observation. During a training phase, a supervised learning algorithm keeps adjusting model parameters to minimize the discrepancy between the target variable and the computed output until a stopping rule is satisfied. This training phase may take a large number of iterations for the model convergence. Once the model is trained, the model is used for predicting target values in a test data set that includes only the feature variables (Bishop, 2007).

In unsupervised learning, no target variable is identified, and all variables can be considered as feature variables. There is no training phase in unsupervised learning

because no target variable exists to “supervise” algorithms. Unsupervised learning methods uncover hidden patterns and structures among all feature variables. Unsupervised learning algorithms build descriptive models rather than predictive models. In the following, market basket analysis, an unsupervised learning method, is explained and explored as to how it can be used to uncover characteristics associated with aberrant test takers who were flagged.

Market basket analysis is used to identify which feature variables are associated with one another. This technique has been utilized in marketing to uncover association rules for quantifying the relationships between two or more features on consumers’ shopping transactions (Larose & Larose, 2014). For example, a grocery store observed that of 100 customers who visited their store one day, 40 bought beer. Furthermore, out of those 40, 10 purchased chips. Then, market basket analysis algorithms create a rule such as “if buy beer, then buy chips {beer → chips}” with a *support* of $10/100 = 10\%$ and a *confidence* of $10/40 = 25\%$ (see Larose & Larose, 2014, for more examples). The support of item is the ratio of the number of the antecedent item transactions to the number of all transactions. The confidence is expressed as the ratio of the transaction number for both the antecedent and consequent items to the number of the antecedent item transactions. The support for the rule {beer → chips} is defined as

$$\text{support}(\text{beer} \rightarrow \text{chips}) = \frac{\text{number of customers bought beer}}{\text{total number of customers}} \quad (22)$$

and the confidence for the rule is defined as

$$\text{confidence}(\text{beer} \rightarrow \text{chips}) = \frac{\text{number of customers bought both beer and chips}}{\text{number of customers bought beer}} \quad (23)$$

Once the grocery store identified which items were purchased together frequently (e.g., beer and chips), based on either high support or high confidence values, the retailer can arrange these items closely so the store can increase its sales on these associated items. There is no target variable in the above example. Researchers are interested in which items are purchased together. The quantifying task is daunting because having k items produces 2^k possible item sets. The Apriori algorithm (e.g., Borgelt, 2012) is designed to efficiently navigate through large item sets looking for sets of items that are strongly associated with each other.

Within the testing context, there are many potential feature variables available (e.g., item responses, response times, demographics, testing center information). As discussed previously, several other models exist to identify aberrant item response or response time patterns. The value of market basket analysis in identifying potential aberrance lies in its ability to discover prevalent characteristics among feature variables other than item response and response time variables.

Data Analysis Using Machine Learning Algorithm

Market basket analysis was conducted using *arules* R package (Hahsler, Buchta, Gruen, Hornik, & Borgelt, 2014). The purpose of the analysis was to identify common characteristics among the flagged test takers. The analysis included only the flagged test takers from the data provider: 46 cases from Form 1 and 48 cases from Form 2. Figure 4.15 provides relative frequencies of prevalent features for each form. The unflagged test

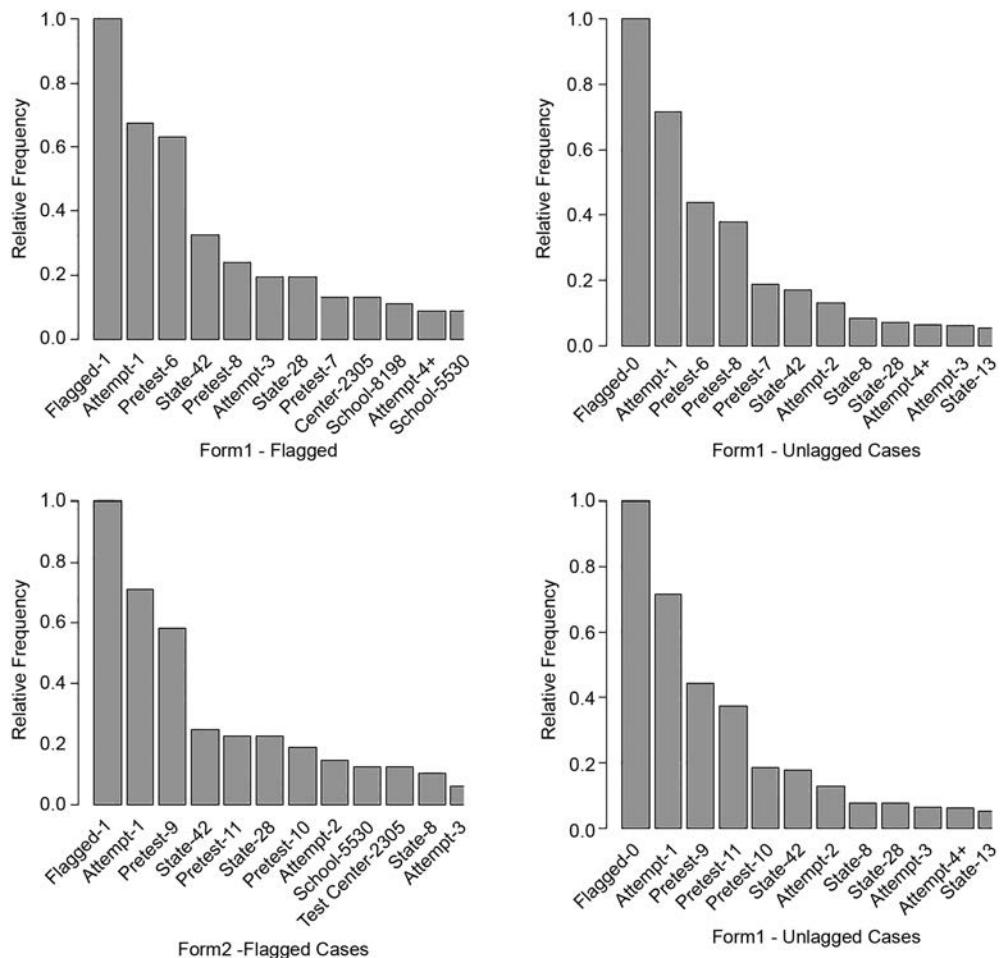


Figure 4.15 Common Features Associated with the Flagged Test Takers

takers were included for comparison purpose. Without this comparison, it was not easy to see that the characteristics associated with the flagged test takers were salient only among the flagged test takers.

It appears that TestCenter-2305, School-8198, and School-5530 were relatively dominant features among the flagged test takers from Form 1, and TestCenter-2305 and School-5530 were from Form 2. It will be worthwhile to investigate why these features are strongly associated with the flagged cases (e.g., the testing center had some security issue or the schools provided any testing materials to their students prior to the exam). Both TestCenter-2305 and School-5530 were associated with the flagged cases across two forms.

Market basket analysis is usually applied to big retail transaction data sets. It is promising to see that the analysis provided valuable information even with a small data set like the testing data. For relatively straightforward datasets, simple frequency analysis could potentially identify the same characteristics. However, as the number of levels in a feature variable (e.g., there are 346 schools in Form 1) and the number of feature variables (e.g., school, testing center, and country) in a dataset increase, it is almost impossible for simple frequency analyses to uncover common characteristics.

CONCLUSIONS

Any person or organization administering high-stakes tests to make important decisions about test takers should be confident that the scores from the test takers are reliable and valid indications of the test takers' abilities on the construct that the test is intended to measure. Because aberrant testing patterns could have a negative impact on the validity of test scores, it is important to identify and investigate these patterns before making any decision. Psychometrics-based and machine learning-based methods were explored to tackle the complex problem of detecting aberrant testing patterns. Some of these methods rely on rather strong underlying assumptions of the data (e.g., unidimensionality, local independence, constant testing speed, following specific statistical distributions). As a result, it is important to check the tenability of the assumptions because the accuracy of the results depends on the degree to which the assumptions are met in the data. In addition, a model-data fit should be verified prior to drawing inferences from the data.

Several person-fit indexes were utilized to identify aberrant item responses. The performances of l_z and l_z^* were very similar. That is, they tend to flag the same test takers. U3 is promising in that it does not require IRT parameter estimates and its performance is on par with the two parametric indexes. All the person-fit indexes identified only a subset of the cases flagged by the data provider. It is unknown how the data provider flagged the test takers except that some of these test takers had item preknowledge. Because there may be some other reasons other than aberrant response patterns for which the data provider flagged, any single person-fit index is not expected to flag all the cases flagged by the data provider.

Two response time models were used to detect aberrant response latencies. Both models produced a nearly identical pattern of chi-squared values with different scales on the y -axis in Figure 4.10 and Figure 4.14. The flagging criterion, when applied to the ERT model, appeared too liberal, whereas the same criterion appeared very conservative in the hierarchical frame model. It was discussed that the process of summing all squares of the residuals may make it difficult to identify aberrant response patterns observed only in a subset of items.

One unsupervised machine learning method, market basket analysis, was used to uncover common characteristics associated with the flagged test takers. The flagged datasets were relatively small (46 cases from Form 1 and 48 cases from Form 2) for data mining, but the method appears promising as a way to summarize the background characteristics that are common among examinees identified with unusual testing patterns. In the example provided here, it was shown how this method can provide useful information on which school and testing center might warrant further investigation.

NOTE

1. A computer-based licensure program provided datasets used in this chapter. The datasets were commonly used across several chapters in this book. Authors' affiliation neither owns nor produces the datasets.

REFERENCES

- Armstrong, R. D., & Shi, M. (2009). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement*, 33(5), 391–410.
- Armstrong, R. D., Stoumbos, Z. G., Kung, M. T., & Shi, M. (2007). On the performance of the l_z person-fit statistic. *Practical Assessment, Research & Evaluation*, 12(16). Retrieved from <http://pareonline.net/getvn.asp?v=12&n=16>

- Bird, C. (1927). The detection of cheating on objective examinations. *School and Society*, 25(635), 261–262.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, A. Birnbaum, & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 425–435). Reading, MA: Addison-Wesley.
- Bishop, C. M. (2007). *Pattern recognition and machine learning*. New York: Springer.
- Borgelt, C. (2012). Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6), 437–456.
- Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association*, 93(443), 910–919.
- Callahan, D. (2004). *The cheating culture: Why more Americans are doing wrong to get ahead*. Orlando, FL: Harcourt.
- Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement*, 15(2), 171–191.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1), 67–86.
- Dwyer, D. J., & Hecht, J. B. (1994). Cheating detection: statistical, legal, and policy implications. Retrieved from <http://files.eric.ed.gov/fulltext/ED382066.pdf>
- Fox, J.-P., Entink, R. K., & van der Linden, W. J. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, 20(7), 1–14.
- Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, 2(4), 235–256.
- Hahsler, M., Buchta, C., Gruen, B., Hornik, K., & Borgelt, C. (2014). arules: Mining association rules and frequent item sets (Version 1.1-6). Retrieved from <http://cran.r-project.org/web/packages/arules/index.html>
- Haney, W. M., & Clarke, M. J. (2006). Cheating on tests : Prevalence, detection, and implications for online testing. In E. M. Anderman & T. B. Murdock (Eds.), *Psychology of academic cheating* (pp. 255–287). Burlington, MA: Elsevier Academic Press.
- International Test Commission. (2013, October 12). ITC guidelines on quality control in scoring, test analysis, and reporting of test scores. Retrieved from www.intestcom.org/upload/sitefiles/qcguidelines.pdf
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. New York, NY: Springer.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16(4), 277–298.
- Kim, D., De Ayala, R. J., Ferdous, A. A., & Nering, M. L. (2011). The comparative performance of conditional independence indices. *Applied Psychological Measurement*, 35(6), 447–471.
- Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: An introduction to data mining*. Hoboken, NJ: John Wiley & Sons.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4(4), 269–290.
- Li, M. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, 21(3), 215–231.
- Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, 39(6), 426–451.
- McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, 27(2), 121–137.
- Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, 9(1), 3–8.
- Meijer, R. R. (2002). Outlier detection in high-stakes certification testing. *Journal of Educational Measurement*, 39(3), 219–233.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2), 107–135.
- Meijer, R. R., & Sotaridona, L. S. (2006). *Detection of advance item knowledge using response times in computer adaptive testing* (LSAC Computerized Testing Report No 03–03). Newtown, PA: Law School Admission Council.
- Meijer, R. R., Tendeiro, J. N., & Wanders, R. B. K. (2014). The use of nonparametric item response theory to explore data quality. In S. P. Reise & Dennis Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 128–155). New York, NY: Routledge.
- Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*. Cambridge, MA: The MIT Press.

- Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research*. The Hague, The Netherlands: Mouton.
- Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, 55(1), 75–106.
- Montgomery, D. C. (2012). *Statistical quality control*. Hoboken, NJ: John Wiley & Sons.
- Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement*, 19(2), 121–129.
- Olson, J. F., & Fremer, J. (2013). *TILSA test security guidebook: Preventing, detecting, and investigating test security irregularities*. Washington, DC: Council of Chief State School Officers.
- Pérez-peña, R. (2012, September 7). Studies show more students cheat, even high achievers. *The New York Times*. Retrieved from www.nytimes.com/2012/09/08/education/studies-show-more-students-cheat-even-high-achievers.html
- Reise, S. R. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, 19(3), 213–229.
- Robitzsch, A. (2015). sirt: Supplementary item response theory models (Version 1.3). Retrieved from <http://cran.r-project.org/web/packages/sirt/index.html>
- Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, 55, 3–38.
- Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden*, 7(22), 131–145.
- Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, 16(2), 149–157.
- Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, 66(3), 331–342.
- Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, 20(4), 331–354.
- Tendeiro, J. N. (2014). PerFit: Person fit (Version 1.2). Retrieved from <http://cran.r-project.org/web/packages/PerFit/index.html>
- Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179–203). New York, NY: Academic Press.
- van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, 13(3), 267–298.
- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2), 181–204.
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3), 287–308.
- van der Linden, W. J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics*, 37(1), 180–199.
- van der Linden, W. J., & Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, 68(2), 251–265.
- van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 195–210.
- van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, 26(2), 199–217.
- van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement*, 26(2), 164–180.
- Wainer, H. (2014). Cheating: Some ways to detect it badly. In N. Kingston & Amy Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 8–20). New York, NY: Routledge.
- Zhang, J. (2007). Conditional covariance theory and detect for polytomous items. *Psychometrika*, 72(1), 69–91.
- Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*, 64(2), 129–152.
- Zhang, J., & Stout, W. (1999b). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*, 64(2), 213–249.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Section IIb

Detecting Preknowledge and
Item Compromise



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

5

DETECTING PREKNOWLEDGE AND ITEM COMPROMISE

Understanding the Status Quo

Carol A. Eckerly

INTRODUCTION

Developments in technology have aided the construction of high-quality assessments, where some examples include the creation of on-demand computer-based testing and computerized adaptive testing. However, at the same time these developments have also facilitated the overexposure of items that can lead to invalid scores. Overexposure of exam content becomes a problem when it leads to some examinees having prior access to the live content before sitting for the exam, a phenomenon known as item preknowledge. Zara (2006) defines item compromise as occurring when an item's performance has changed over time, and that change is due to "its content having been distributed beyond its defined valid usage boundaries" (p. 2). Therefore, item exposure is a necessary (barring an internal breach) but not sufficient condition for item compromise to occur.

Examinees can gain preknowledge from a variety of different sources of item compromise. For fixed form exams that have continuous testing windows, prior to sitting for the exam, examinees may discuss exam content with colleagues or classmates who have already taken the exam. On a larger scale, educators may ask students to memorize items from an exam to share them with future classes. And for potentially much broader impact on the validity of scores, examinees who steal items could post them to a brain dump website or create a rogue review course in an attempt to profit from the theft.

There have been several high-profile cases involving item compromise that can serve to remind us of the potential severity of consequences this type of security threat can have on exam programs. One example occurred in 2010, when 137 doctors were suspended by the American Board of Internal Medicine (ABIM) for sharing exam content with an exam prep company. This was the first of two recent known security issues for the American Board of Medical Specialties (ABMS), the umbrella group for 24 medical boards. The second security issue involved the American Board of Radiology, where

the widespread use of “recalls” (i.e., compromised items) by radiology residents at The San Antonio Uniformed Services Health Education Consortium (SAUSHEC) was uncovered. The use of the recalls, which were made available on a website for radiology residents and a shared military server, was encouraged by the program director (Zamost, Griffin, & Ansari, 2012). When CNN became aware of this story, they interviewed many radiologists to determine if this use of recalls was standard practice or limited to SAUSHEC. CNN investigators found that radiologists readily admitted that use of recalls was widespread. Several radiologists were even quoted with their justifications as to why using the recalls did not constitute cheating. It is truly alarming that despite laws in place to protect exam content, user agreements to inform candidates that disclosing exam content is considered cheating, and publicized incidents of doctors being suspended for sharing live exam content, some doctors were still willing to go on record stating that this behavior is not wrong. This example should highlight the severity of threats to score validity due to item compromise and remind test developers of the challenges they face regarding item compromise.

Another high-profile example of exam compromise occurred in 2002, when Educational Testing Service (ETS) discovered that many students in China, Taiwan, and South Korea were benefiting from brain dump sites showing live Graduate Record Exam (GRE) content from computerized adaptive tests. Compromise was so widespread that average scores increased by 100 points (out of a possible 1,600 points) in China and 50 points in South Korea (Kyle, 2002). In response, ETS suspended computerized testing in China, Taiwan, and South Korea, and only offered paper and pencil versions until 2011. However, students in these countries could still take the computerized exams by traveling to another country to sit for the exam. One Chinese student who used this strategy after studying the compromised items stated that her extra travel cost was “money well-spent” (Hornby, 2011).

In scandals revealing widespread compromise such as these, it is not only the cheating students who are affected by their cheating behaviors. The validity of the scores of students who legitimately prepared for an exam and honestly received a high score may be questioned, as well. Furthermore, when some examinees respond to items using their true abilities as well as preknowledge of some subset of items, all item and person parameter estimates will be contaminated. As the percentage of compromised items and the percentage of examinees with preknowledge increases, the magnitude of contamination of parameter estimates will increase as well, with an increasingly negative impact for honest examinees.

DETECTION

Given the severe consequences of item compromise to testing programs, both measures to prevent item compromise and detect it when it occurs are necessary. While efforts to prevent item compromise are arguably more important for the well-being of a testing program, I will only focus on detection methods here. For readers interested in prevention methods, see Fitzgerald & Mulkey (2013), Foster (2013), and Scicchitano & Mead (2013). Detection methods to deal with the problem of item compromise and examinee preknowledge can be organized into four categories:

1. Methods to identify examinees at the individual level who may have benefitted from preknowledge of compromised items
2. Methods to identify items that may have been compromised

3. Methods to identify both individual examinees who may have benefitted from item preknowledge and the items to which they had prior exposure
4. Methods to identify groups of examinees that may have benefitted from item preknowledge.

Given the variety of ways in which the methods discussed in this chapter work and the aspects that they attend to, designing a single simulation that would compare them all is impractical. Furthermore, because all of the methods discussed in this chapter are already established, they have already been subject to simulation studies and other tests that demonstrated their efficacy. Many of the approaches discussed in this chapter are also described and extended in other chapters of this volume, hence simulations and/or applications to one of the common data sets for those methods are provided there. However, because the scale-purified deterministic gated item response theory model (scale-purified DGM) (Eckerly, Babcock, & Wollack, 2015) is recent and promising method that is not discussed elsewhere in this volume, this method is applied to the common licensure data set and compared with the deterministic gated item response theory model (DGM) (Shu, Leucht, & Henson, 2013).

In the discussion of the methods in this chapter, emphasis is placed on describing the nature of threats due to preknowledge of live exam content and highlighting differences and similarities between the methods, both in terms of purpose and methodology. I also offer a set of general guidelines to consider when evaluating or using any method to detect item preknowledge and provide suggestions for the direction of future research.

Evaluating Methods for the Detection of Preknowledge

Before discussing individual methods, it is first helpful to spend some time thinking about what properties an optimal detection method would have. If it was possible, we would want to detect all of the examinees with preknowledge or compromised items (i.e., 0% false negative rate) and concurrently avoid any false positive detections. While these goals are unrealistic under most circumstances, we should keep in mind that we hope to get as close as possible to these ideals with any detection method. Furthermore, we should not only be interested in the two diagnostic measures that are most commonly shown in the literature (i.e., some form of false positive and false negative rate), but we should also be interested in what Skorupski & Wainer (this volume) refer to as the Posterior Probability of Cheating (PPoC). The PPoC is the probability that an examinee is a cheater given that he or she is flagged (or, in the context of identifying compromised items, the probability that an item is compromised given that it is flagged). This measure can also be referred to as the true detection rate. True detection rates are important to consider in addition to false positive and false negative rates because they remind us that we are really interested in knowing how to interpret the results of an individual flag. In instances with very low rates of compromise for either individuals or items, it would be misleading to only consider a false positive rate diagnostic because it is likely that the number of false positive flags will overwhelm flags resulting from compromise. This is especially important to consider given that it usually would not be possible to know a priori what the base rate of individuals with preknowledge or compromised items is, and for constant measures of false positive rate and false negative rate, true detection rate will vary widely for different base rates. For a more thorough discussion, see Skorupski and Wainer (this volume).

It is a general rule that the detection power of a method will vary based on a variety of factors, including but not limited to examinee ability, base rate of examinee preknowledge, and compromised item difficulty. For example, if we are interested in detecting individual examinees who have preknowledge of a subset of items, we can imagine that it would be almost impossible to detect high-ability examinees that have had preknowledge to a subset of very easy items in absence of some nonstatistical evidence. We are much more likely to detect low-ability examinees who have had preknowledge to a subset of very difficult items. We should not be too concerned if the power of a method varies depending on factors such as these, as long as the method is reasonably good at detecting the types of aberrances that most severely jeopardize the validity of the exam scores or constitute a threat to public safety (e.g., a medical doctor who is not competent to practice but passes his or her board exam due to help from item preknowledge). However, we should be very concerned if the false positive rate varies considerably and/or is higher than the nominal level, depending on the level of a factor that is unknowable in practice (such as base rate of examinees with preknowledge in the testing population). If an examinee or item is flagged by a detection method, we hope to be able to make an accurate probabilistic statement about that flag.

The earliest efforts for detecting preknowledge involve the use of person-fit statistics (McLeod & Lewis, 1999; Meijer & Sijtsma, 1995; van Krimpen-Stoop & Meijer, 2000). A person-fit statistic detects item score patterns that are unlikely given the IRT model that is employed. Given this definition, we can imagine that response data of examinees who have benefitted from preknowledge might be flagged using a person-fit statistic. However, there are many other legitimate testing behaviors that could also be flagged by a person-fit statistic, such as misfit due to random response behavior, non-invariant ability due to carelessness or fumbling, or violations of local independence (van Krimpen-Stoop & Meijer, 2000). Therefore, the use of person-fit statistics alone should not be considered evidence that an examinee had preknowledge.

METHODS TO IDENTIFY INDIVIDUALS WHO HAVE BENEFITTED FROM ITEM PREKNOWLEDGE

The Deterministic Gated Item Response Theory Model

There are a variety of circumstances in which a testing organization may learn that a subset of live items has been compromised. Some examples include the organization becoming aware of a brain dump site distributing live exam content, or the organization obtaining a list of live items shared by an educator. In circumstances such as these where the organization is fairly certain that a specific subset of items has been compromised, the Deterministic Gated Item Response Theory Model (DGM) (Shu et al., 2013) can be employed to flag individuals that may have benefitted from preknowledge of the specified items.

The DGM is defined by Equations 1 through 5:

$$\begin{aligned} P(U_{ij} = 1 | \theta_{tj}, \theta_{cj}, b_i, T_j, I_i) \\ = P(U_{ij} = 1 | \theta_{tj})^{1-T_j} * \left[(1 - I_i) * P(U_{ij} = 1 | \theta_{tj}) + I_i * P(U_{ij} = 1 | \theta_{cj}) \right]^{T_j} \end{aligned} \quad (1)$$

$$P(U_{ij} = 1 | \theta_{tj}, b_i) = \frac{\exp(\theta_{tj} - b_i)}{1 + \exp(\theta_{tj} - b_i)} \quad (2)$$

$$P(U_{ij} = 1 | \theta_{cj}, b_i) = \frac{\exp(\theta_{cj} - b_i)}{1 + \exp(\theta_{cj} - b_i)} \quad (3)$$

$$\sum b_i = 0 \quad (4)$$

$$T_j = 1, \text{ when } \theta_{cj} < \theta_{cj}, \quad (5)$$

where θ_{tj} is the j th examinee's true ability, θ_{cj} is the j th examinee's cheating ability (which is estimated based on the examinee's performance on the items specified as compromised), and b_i is the item difficulty for item i . I_i is the gating mechanism indicating whether item i is assumed by the user to be secure ($I_i = 0$) or compromised ($I_i = 1$). From Equation 1, if an item is secure, then responses of both examinees with and without preknowledge for that item are based on examinees' true abilities. However, if an item is compromised, the responses of examinees who did not have preknowledge of the item are still based on their true abilities, but responses of examinees who did have preknowledge of the item are based on their cheating abilities.

By using Markov Chain Monte Carlo (MCMC) estimation, for each iteration, the model classifies each examinee into one of two latent classes—one for those performing better on the items specified by the user as compromised ($T_j = 1$), and one for examinees performing equivalently or better on the set of items specified by the user as secure ($T_j = 0$). Across all post burn-in iterations of the MCMC run, the proportion of posterior samples in which examinee j is assigned to the preknowledge class is labeled T_j^* . Examinees whose value of T_j^* exceeds some user-specified threshold (e.g., 0.95) are categorized as individuals with preknowledge, whereas those with T_j^* values below the threshold are categorized as examinees without preknowledge. The usual trade-off between false positives and false negatives can be shown by varying the user specified threshold (e.g., from 0.95 to 0.99).

Shu et al.'s (2013) simulation study showed the sensitivity (i.e., one minus false negative rate) and specificity (i.e., one minus false positive rate) for varying proportions of compromised items, percentages of examinees with preknowledge, and levels of cheating effectiveness for a 40-item test. For the simulation study, examinees with preknowledge were simulated based on their true abilities, where lower ability examinees were more likely to be simulated as having preknowledge than high-ability examinees. Results of the simulations showed specificity varied across conditions from 1.00 to 0.94, with the highest levels being associated with high percentages of compromised items and high percentages of cheaters. Sensitivity across conditions was less consistent, ranging from 0.14 to 0.82, with the highest levels being associated with equal proportions of compromised and secure items and low percentages of cheating examinees.

Although the simulations conducted by Shu et al. (2013) showed the promise of the DGM in detecting individuals with preknowledge, there are several limitations that highlight the need for future research and refinement of the model. First, Shu et al. only simulated fairly large percentages of compromised items (30%, 50%, and 70%). These percentages may very well be realistic for some testing programs. For example, IT certification programs have been known to have very high amounts of compromise, both in terms of percentage of compromised items and percentage of examinees having preknowledge of those items (Maynes, 2009). But if it was a general rule for testing programs to have such high levels of compromise, we should be wary of the validity of any test score. These percentages of compromise are truly a worst case scenario;

therefore, it is also important to know how models perform with lower amounts of compromise as well as in the absence of any preknowledge.

Furthermore, Shu et al. (2013) did not address how the model works when practitioners have imperfect information as to which items are compromised. It is possible for items that are truly compromised to be misspecified by the practitioner as secure, or vice versa. The former is probably most likely. As an example, a brain dump site may be found that has some subset of compromised items that the user inputs into the model. In this example, the user has very strong evidence that the items specified as compromised truly are compromised, but that subset of items is not necessarily exhaustive. However, the latter can effectively occur from a detection standpoint if an item has been compromised but very few individuals have had prior access to it.

Last, in Shu et al.'s (2013) simulations, the θ distribution of the examinees with preknowledge was positively skewed, which biases the study towards favorable results. When examinees with preknowledge are predominantly low ability, they have more to gain from the preknowledge, and thus they will be easier to detect.

Eckerly and Wollack (2013) conducted a simulation study to show how the DGM performs when data are simulated under different conditions to address these limitations. Examinees with preknowledge were randomly selected independent of their ability, smaller percentages of compromised items were simulated, and various levels of user misspecification of item compromise status were simulated. In these differing conditions, model performance was much poorer. False positive rates increased as high as 31%, showing that there are some circumstances in which model results are grossly inaccurate. In terms of false positive rate, the model performed least favorably with low percentages of compromised items, misspecification of compromised items as secure, and a short test length. False positive rate varied significantly depending on the size of the subset of items specified as compromised, regardless of whether the items were truly compromised, because the model performs best under conditions where the sets of items used to estimate θ_{tj} and θ_{cj} are of sufficient size to accurately estimate both θ_{cj} and θ_{tj} .

This study highlights the need for practitioners to develop some intuition for whether a flagged examinee likely benefited from item preknowledge or whether the flag is errant. One strategy to aid in this endeavor for the DGM is to conduct simulations to find the expected false positive rate for a given number of compromised items when *no examinees have preknowledge* (Eckerly et al., 2015). For example, if a user has evidence to believe that 30 items on a 200 item exam have been compromised, the user can simulate data to mirror his or her exam under conditions of no compromise. This might include simulating item and person parameters based on historical distributions for the given exam size, then randomly selecting 30 items for each replication of the simulation. If the practitioner then runs the DGM using the actual response data, specifying the 30 items thought to be compromised as such, that result can be compared to the expected false positive rate when no examinees have preknowledge. If these numbers are similar, then the practitioner may conclude that the flags are likely errant flags.

Scale Purified Deterministic Gated Item Response Theory Model

Eckerly et al. (2015) described how the DGM methodology leads to biased item and person parameter estimates, even when there is no user misspecification of the set of compromised items, because misclassifications of examinees at each iteration of the MCMC run affect the item difficulty estimates. Eckerly et al. described two sources

of bias in item difficulty estimation, both of which have the potential to severely jeopardize the model's ability to accurately classify examinees into the preknowledge class. The first source of bias results from the model's omission of response data from a portion of examinees in the estimation of item difficulties, which leads to upward bias in the item difficulty estimates. This upward bias in the difficulty estimates for compromised items results in the appearance of examinees doing better than expected on those items (i.e., θ_{cj} estimates for all examinees will be spuriously high). Consequently, the DGM will report an inflated false positive rate of examinees flagged as suspicious.

The second source of bias was addressed by Shu et al. (2013), where the true ability scale drifts to resemble a composite of both true ability and score gain due to item preknowledge, leading to downward bias in the item difficulty estimates of compromised items. The magnitude of the bias from this source depends on the percentage of examinees with preknowledge in the population. These biased estimates in turn affect false positive rates, power, and true detection rates.

To reduce bias in item and person parameter estimates, a modification of the DGM methodology was proposed by Eckerly et al. to purify the estimates. The authors introduced an iterative scale purification procedure that includes the following steps:

1. Using all of the response data, estimate item difficulty parameters using the one-parameter logistic model (Rasch, 1960).
2. Rather than estimating the item difficulty parameters in the DGM, fix them to the parameter estimates from Step 1, then run the DGM.
3. Remove the response data from the flagged examinees and reestimate item difficulty parameters using the one-parameter logistic model.
4. Using the purified item difficulty estimates from Step 3, run the DGM again with the response data of all examinees.

By fixing the item difficulties at each iteration, the item difficulty estimation bias resulting from the omission of response data of honest examinees is eliminated, and by removing the response data of examinees flagged in Step 2 and reestimating item difficulty parameters to be used in the next iteration of the DGM (i.e., Step 4), the bias resulting from the scale drift is minimized. In comparison to the original DGM methodology, Eckerly et al.'s (2015) simulation showed that false positive rates dramatically decreased and true detection rates increased under most circumstances using the scale-purified DGM. In instances where there was a low base rate of examinees benefitting from preknowledge, the scale purification steps (i.e., steps 3 and 4 of the procedure) did not add much benefit beyond using the DGM fixed contaminated item parameter estimates (i.e., only completing steps 1 and 2 of the procedure). For high base rates of examinees benefitting from item preknowledge, the scale purification steps significantly decreased false negative rates in comparison to their absence, although true detection rates remained stable.

As mentioned earlier in this chapter, both the original DGM methodology and the scale-purified DGM are applied to the licensure common data set that consists of two forms which each contain 170 scored items. Pilot items are excluded from the analysis. Data were available for 1,636 candidates in Form 1 and for 1,644 candidates in Form 2. Form 1 has 63 items identified by the licensure organization as compromised, and Form 2 has 61 items identified as compromised. These classifications (i.e., an item is compromised or it is not) are used as inputs into both the DGM and the scale-purified DGM,

although the criteria which was used to determine these classifications is unknown to the author.

As Eckerly et al. (2015) suggest, it is helpful to first conduct simulations to understand how these methods perform under conditions without any compromise. This step is necessary because expected false positive rate in conditions where no examinees have benefitted from preknowledge varies significantly depending on the size of the subset of compromised items. For these simulations, which are referred to as null conditions, the numbers of items specified as compromised corresponded to the number of compromised items from the two exam forms in the real data example (i.e., 61 and 63). For each condition, 15 replications were run where true ability and item difficulty parameters were both generated from *normal* (0,1) distributions. In each replication, a different random subset of items was misspecified as compromised, for purposes of estimating the DGM. For the scale-purified DGM, intermediate T_j^* threshold = 0.95. Flagging results including the mean and standard deviation of the percentage of flagged examinees across the 15 replications are shown in Table 5.1.

As T_j^* threshold increases from 0.95 to 0.99, the percentage of flagged examinees decreases for both the DGM and the scale-purified DGM (as we would expect), and flagging rates for both methods are below the nominal level. The expected false positive rate for the null conditions is low due to the fairly high number of items specified as compromised in both forms. When T_j^* threshold = 0.95, the expected false positive rate for the null conditions is lower for the scale-purified DGM in comparison to the DGM, but results are comparable for the two methods when T_j^* threshold = 0.99.

The flagging results of these null conditions can be compared to the flagging results from both forms of the real data, which are shown in Table 5.2.

Table 5.1 % Flagged Individuals—Simulated Null Conditions

		Flagging Threshold	Form 1	Form 2
DGM	mean	$T_j^* = 0.95$	0.0223	0.0207
	<i>SD</i>		0.0048	0.0033
	mean	$T_j^* = 0.99$	0.0064	0.0067
	<i>SD</i>		0.0018	0.0018
Scale-Purified DGM	mean	$T_j^* = 0.95$	0.0142	0.0157
	<i>SD</i>		0.0029	0.0023
	mean	$T_j^* = 0.99$	0.0057	0.0073
	<i>SD</i>		0.0010	0.0021

Table 5.2 % Flagged Individuals—Real Data

	Flagging Threshold	Form 1	Form 2
DGM	$T_j^* = 0.95$	0.1051	0.2041
	$T_j^* = 0.99$	0.0605	0.1381
Scale-Purified DGM	$T_j^* = 0.95$	0.0562	0.1045
	$T_j^* = 0.99$	0.0275	0.0839

For the flagging criteria shown (as established by the level of T_j^*), the DGM and the scale-purified DGM show much higher percentages of flagged examinees than the corresponding null conditions. This provides some evidence that a portion of examinees flagged from the real data truly benefitted from preknowledge of the subset of compromised items. However, we should not be too hasty jumping to this conclusion. Results from Eckerly et al.'s (2015) similar analysis comparing real data flagging and null condition flagging are relevant here. For reasons unique to their real data, the authors concluded that the vast majority of flags from the real data analysis were likely to be errant flags (i.e., false positives). However, despite this, in comparing the results for the null conditions vs. the real data conditions, the real data conditions exhibited much higher percentages of flagged examinees than the simulated null conditions. These differences were greater when the subset of compromised items was smaller. Drawing from the results of Eckerly et al. and the results shown here, it seems reasonable to conclude that there is some nontrivial percentage of examinees who benefitted from preknowledge, but the flagged individuals most likely also consist of some nonnegligible and not easily quantifiable percentage of false positives.

Interestingly, Form 2 resulted in higher percentages of flagged examinees than Form 1 for both the DGM and the scale-purified DGM. It seems reasonable in this case to conclude that many more examinees were benefitting from preknowledge in Form 2 than Form 1 because the size of the subset of compromised items is very similar for the two forms, so the number of items specified as compromised is large enough to have a fairly accurate estimate of θ_{cj} . Also, Forms 1 and 2 are more directly comparable to each other than to simulations because they both consist of real data presumably from a similar population of examinees.

In comparing the scale-purified DGM to the DGM, we can also see that the percentage of flagged examinees for a given flagging criteria decreases when using the scale-purified DGM in comparison to the original DGM methodology for both exam forms. This is an indication that the scale-purified DGM produced both fewer false positives and fewer true positives than the original DGM for both Form 1 and Form 2. Based on results from Eckerly et al.'s (2015) simulations, it is likely that the decrease in false positives overwhelmed the decrease in true positives, leading to a higher true detection rate (i.e., percentage of examinees who truly benefited from preknowledge out of those who were flagged by the method).

Although these results do not give conclusive evidence that any individual examinee had preknowledge, they do give some indication of how widespread the compromise was, and they could serve as a starting point for further investigations. The results could be used to search for patterns among flagged examinees and to inform discussion about the possibility of retiring certain items. However, without the full context of the breach, it is difficult to make any concrete recommendations here.

Trojan Horse Items

The effects of item compromise are probably the most severe in instances when an actual exam file along with the key is stolen. Maynes (2009) introduced a methodology which makes use of a set of miskeyed, unscored items called Trojan Horse items. The idea behind this methodology is that when the exam and the key are stolen and posted to a brain-dump site, beneficiaries of the stolen content will attempt to memorize all of the items and their answers. High-ability examinees may notice the error in the key, but lower ability examinees will likely respond to the item incorrectly

based on memorization of the key. Even examinees of higher ability may put more faith in the incorrectly keyed response than in their own abilities. Thus, it is expected that examinees who have had access to the exam questions and the key will do very well on the operational items and very poorly on the Trojan Horse items (because they will respond based on the incorrect key). For each examinee, it is possible to calculate the probability that he or she answered the given number of Trojan Horse items incorrectly, given his or her estimated ability on the operational items. If this probability is sufficiently low, the examinee can be flagged as an individual with preknowledge.

FLOR Log Odds Ratio Index and Differential Person Functioning

Both McLeod, Lewis, and Thissen (2003) and Smith and Davis-Becker (2011) have introduced methods to flag individuals who have benefitted from preknowledge that can be extended to use information from person flagging to inform methods for item flagging as well. McLeod et al. introduced the FLOR log odds ratio index, and Smith and Davis-Becker used Differential Person Functioning (DPF) statistics for person flagging. Because these methods were extended in future studies to include item flagging, they are discussed in more detail in the section of this chapter which describes methods to identify both examinees that have benefitted from item preknowledge and compromised items.

METHODS TO IDENTIFY COMPROMISED ITEMS

Moving Averages

Han (2003) introduced the use of moving averages to help detect items that may have been compromised. When an item is compromised, it is likely that more examinees will answer that item correctly than if the item had been secure, leading the item to appear easier than it actually is. While monitoring changes in item difficulty across time could provide valuable information to search for this pattern in the presence of item preknowledge, Han showed how using moving averages rather than simple averages can more effectively alert practitioners if compromised items drift to become easier. Rather than using item difficulty estimates for the analysis, Han used “*p*-values” (the proportions of correct response for each item).

A moving average is calculated by selecting a window size, k , for analysis of response data of n examinees. For a window size of $k = 100$, the sequence of moving *p*-values for item i (which is denoted as p) is:

$$(p_{100}, p_{101}, \dots, p_{n-100})$$

$$\text{where } p_{100} = \frac{1}{k}(u_{i,1} + u_{i,2} + \dots + u_{i,100}),$$

$$p_{101} = \frac{1}{k}(u_{i,2} + u_{i,3} + \dots + u_{i,101})$$

$$p_{102} = \frac{1}{k}(u_{i,3} + u_{i,4} + \dots + u_{i,102})$$

...

$$p_{n-100} = \frac{1}{k}(u_{i,n-99} + u_{i,n-98} + \dots + u_{i,n}).$$

To illustrate the purpose of monitoring moving averages of p -values rather than simply monitoring average p -value, it is helpful to first visualize simple average p -values across time. When an item is first administered, the sample size of response data for that item is low. As more examinees are exposed to the item over time, the sample size continually increases, as does the stability of the estimated p -value. Thus, after the item has been in operation for a long time, new observations have less of an influence on the overall average than beginning observations did.

In contrast, a moving average shows how the average p -value changes across time for a predetermined, fixed sample size. By examining moving averages rather than simple averages across time, it is easier for practitioners to pinpoint exactly when the average p -value started changing. Thus, items that have potentially been compromised at some point within a testing window can be flagged as soon as possible.

The use of moving averages of p -value relies on the restrictive assumption that the distribution of examinee ability is stationary over time. This assumption is not realistic in a variety of contexts. Han and Hambleton (2004) expanded the use of moving averages by comparing the performance of three statistics in detecting item compromise: (1) Moving p -values, (2) moving averages of item residuals, and (3) moving averages of standardized item residuals. The authors evaluated the performance of these three statistics under a variety of conditions, most notably varying whether the assumption of a stationary ability distribution across time is met. The authors found that the moving p -values approach performed very poorly when the assumption of a stationary ability distribution across time was not met (flagging almost every item, even if it was not compromised), but the other two statistics performed fairly well in these cases. The moving averages of item residuals and standardized item residuals performed well except in cases where very small levels of compromise were simulated and items were not very difficult (which is to be expected).

It is helpful to examine several aspects of this study to extend more broadly when thinking about item compromise or examinee preknowledge detection. First, the authors discussed simulating an empirical sampling distribution of statistics to get an idea about how much variation to expect when there is *no compromise*. Although this is an important step in any operational testing context, it is often overlooked in test security research. Because the presence of examinee responses using item preknowledge will most often contaminate parameter estimates used in cheating detection methods, understanding flagging behavior in the absence of compromise is extremely important.

Second, the authors simulated cheating behavior to represent two very different types of item compromise. In the first type of item compromise, examinees with preknowledge responded correctly to compromised items with probability of 1. This type of response behavior is likely consistent with an examinee who has seen an exact copy of the item with the correct key and has carefully memorized the item. In the second type of item compromise, examinees with preknowledge were simulated to have an increased probability of correctly answering the item, which is only slightly higher than the probability based on their true abilities. For example, an

examinee whose probability of correct response was 0.25 based on true ability for one example item was simulated to answer correctly to that item with probability of 0.44. This type of response behavior is likely consistent with some general information about the item being shared, maybe through conversation with an examinee who has previously taken the exam but wasn't attempting to steal content. Although this type of behavior still represents item compromise, its consequences are likely not as severe.

The authors' choice to simulate response data reflecting item compromise in two different ways calls to our attention that we do not actually know how item preknowledge manifests itself in response patterns. Simulating responses based on item preknowledge using various plausible methods is a useful strategy to help us have a better understanding of how a preknowledge detection method functions.

Log Odds Ratio Statistic

Obregon (2013) presented a method for detecting items that are likely to be compromised which is similar to McLeod and Schnipke's (2006) FLOR_i index (discussed in detail in the upcoming section). Whereas McLeod and Schnipke examined a log odds ratio test in a CAT environment using a three parameter logistic IRT model to detect both compromised items and examinees with preknowledge, Obregon's method used a Rasch framework and only examines the log odds ratio statistic at the item level. The log odds ratio is calculated for each item, and the results indicate how likely it is that each item is compromised given the examinees' response patterns relative to a user specified prior probability that the item is compromised. As an illustration, an index value of 1 calculated for an item would indicate that the probability of compromise given the examinee response data is 10 times greater than the user specified prior probability of compromise. This method is primarily used to flag certain items as suspicious to trigger further investigation.

The log odds ratio index uses Bayes' theorem to calculate the posterior probability that item i is compromised given its observed response vector, X_i :

$$p(c|X_i) = \frac{p(c)p(X_i|c)}{p(c)p(X_i|c) + p(\bar{c})p(X_i|\bar{c})}, \quad (6)$$

where $p(c)$ is the prior probability that item i is compromised, $p(\bar{c})$ is the prior probability that item i is not compromised (i.e., $1 - p(c)$), $p(X_i|c)$ is likelihood of observing response vector X_i , given item i is compromised, and $p(X_i|\bar{c})$ is likelihood of observing response vector X_i , given item i is not compromised.

Because the Rasch model (which assumes local independence) is used to estimate item and person parameters, the above equation can be written as:

$$p(c|X_i) = \frac{p(c)\prod_{j=1}^n p(x_{ji}|c, \hat{\theta}_j)}{p(c)\prod_{j=1}^n p(x_{ji}|c, \hat{\theta}_j) + p(\bar{c})\prod_{j=1}^n p(x_{ji}|\bar{c}, \hat{\theta}_j)} \quad (7)$$

where $p(x_{ji}|c, \hat{\theta}_j)$ is the probability of observing response x_{ji} by person j to item i given that the item is compromised, and $p(x_{ji}|\bar{c}, \hat{\theta}_j)$ is the probability of observing response x_{ji} by person j to item i given that the item is not compromised.

The above equation is evaluated using the conditional item response functions:

$$p(x_{ji}|\bar{c}, \hat{\theta}_j) = \frac{\exp[x_{ji}(\hat{\theta}_j - b_i)]}{1 + \exp(\hat{\theta}_j - b_i)} \quad (8)$$

$$p(x_{ji}|c, \hat{\theta}_j) = p(E_j) \left[\frac{\exp[x_{ji}(\hat{\theta}_j + \hat{D} - b_i)]}{1 + \exp(\hat{\theta}_j + \hat{D} - b_i)} \right] + [1 - p(E_j)] \left[\frac{\exp[x_{ji}(\hat{\theta}_j - b_i)]}{1 + \exp(\hat{\theta}_j - b_i)} \right] \quad (9)$$

where $\hat{p}(E_j)$ = estimated probability that examinee j has preknowledge, and $\hat{p}(E_j)$ = estimated logit difficulty shift, which is used to model the expected decrease in item difficulty in the presence of item compromise.

From (7), using the value of $p(c|X_i)$ calculated using (8) and (9), with the user-specified values of $p(c)$, $\hat{p}(E_j)$, and \hat{D} , the log odds ratio statistic is calculated as:

$$l_o = \log_{10} \left[\frac{p(c|X_i)/[1-p(c|X_i)]}{p(c)/[1-p(c)]} \right] \quad (10)$$

Obregon (2013) conducted a simulation study that illustrated the promise of using the log odds ratio index to identify items that may have been compromised. In his study, Obregon varied the percentage of examinees with preknowledge in the population (5%, 10%, and 20%) and the estimated logit shift (1.0, 2.0, and 3.0), keeping the estimated base rate of examinees with preknowledge at 0.05 regardless of the actual percentage in the population. Data were simulated using the logit shift of 2.0 for all conditions (so we are able to see how misspecification of this logit shift by one logit in either direction affects results). Across all simulated conditions, including some conditions showing misspecifications of the logit shift and the base rate of examinee preknowledge, false positives rates using a flagging criteria of 1.0 were all below 0.4%. Misspecification of the logit shift by either one logit in the positive direction or one logit in the negative direction had no effect on the false positive rate. Hit rate varied between 18% and 70%. Conditions where the actual rate of examinees with preknowledge was the lowest (0.05) showed the lowest power, and power increased as the actual rate of examinees with preknowledge increased. Misspecification of the logit shift either by one logit in the positive direction or one logit in the negative direction had very little influence on the hit rate.

Obregon (2013) assumed that examinees with lower ability are more likely to have item preknowledge than examinees with higher ability. This assumption was reflected both in how examinees with preknowledge were selected and in the detection algorithm itself. Obregon varied the user input $\hat{p}(E_j)$ in the calculation of l_o to reflect the assumption that lower ability candidates are more likely to have item preknowledge. So, even though the overall estimated rate of examinees with preknowledge was a constant value of 0.05, $\hat{p}(E_j)$ varied, conditional on estimated examinee ability. Given that Obregon simulated item responses for examinees with preknowledge based on this assumption, then used the same scheme in the detection method itself, results were likely biased towards favorable outcomes. Even if the assumption that lower ability candidates are more likely to have preknowledge is justified, it is almost certain that

the user could not select $\hat{p}(E_j)$ values at the level of the individual examinee that accurately reflected the true population values, as was assumed in this study.

To address this and other questions, Eckerly et al. (2015) extended Obregon's study by conducting additional simulations and analyzing real data from a national medical imaging certification program. The simulations were designed to further study the behavior of log odds ratio statistics under conditions that varied the amount of item compromise (including conditions with no compromise), type of user misspecification of the user inputs, and two factors unrelated to item compromise (i.e., sample size and the distribution of examinee ability). These noncompromise factors were chosen based on characteristics of the real data.

Eckerly et al. (2015) found that the user input that had the most dramatic effect on the magnitude of the l_o statistics (i.e., $\hat{p}(E_j)$) actually had very little practical impact on the conclusions drawn. The authors illustrated that there is actually a systematic relationship between the set of odds ratio statistics calculated for each item on a given form with a given constant $\hat{p}(E_j)$ across examinees and any other set of odds ratio statistics calculated from the same response data using a different constant $\hat{p}(E_j)$ across examinees. Namely, the second set of odds ratio statistics is almost a perfect linear combination of the first set. Thus, it seems practical for users to choose any constant $\hat{p}(E_j)$ and develop flagging criteria specific to that level of $\hat{p}(E_j)$.

The main conclusion drawn from Eckerly et al.'s (2015) paper was that both changes in user inputs and factors unrelated to item compromise can lead the distribution of odds ratio statistics to resemble what would have occurred in the presence of item compromise (i.e., the variance of the distribution of odds ratio statistics increases in the presence of item compromise). Thus, it is imperative for practitioners to consider other noncompromise hypotheses when the results are consistent with what might be expected if compromise had occurred.

METHODS TO DETECT GROUPS OF EXAMINEES

Detection of Collusion Using Kullback-Leibler Divergence

Belov (2012, 2013, & this volume) introduced a multistage approach to identify individuals who may have benefited from preknowledge within some predefined grouping mechanism, such as testing center or educational program. This method could be used, for example, if we want to flag educational programs (the grouping mechanism) and individuals within these programs that show anomalies when analyzing data from a licensure exam program. Flagged educational programs could be the result of a program director facilitating the sharing of compromised items or a collusion ring involving current and previous students of the program.

Different variations of Belov's method can be used in three separate cases: (1) each group of examinees with preknowledge has had access to the same subset of compromised items, and this subset is known; (2) each group of examinees with preknowledge has access to a subset of items which may be different from other groups with preknowledge, and these subsets are known; and (3) there are no known sets of compromised items. Belov's method is the only method to date that explicitly attempts to detect different subsets of compromised items for different groups of examinees, which is likely the most realistic problem that practitioners face. For detailed simulation results and application to the common data set, see Belov (this volume).

Detection of Collusion Using Cluster Analysis

Wollack and Maynes (this volume) introduced a methodology using both similarity indices and cluster analysis to detect groups of examinees whose responses exhibit unusual similarity. This method is similar in purpose to the method described above using the Kullback-Leibler Divergence (Belov, 2013), but it does not require a grouping mechanism to be applied a priori. For a full discussion of the method and application to the common data set, see Wollack and Maynes.

Detection of Collusion Using Factor Analysis

Zhang, Searcy, & Horn (2011) presented a method that is a sort of hybrid of the Belov (2012, 2013, & this volume) and Wollack & Maynes (this volume) methodologies described above. All three of these methodologies have a similar goal of identifying aberrant groups and individuals within those groups. Like Belov, Zhang et al. identify aberrant individuals using person-fit statistics, and like Wollack and Maynes, a grouping mechanism does not need to be defined a priori. The first step in this method is choosing a person-fit index along with a threshold that would indicate an individual's response pattern is aberrant, assuming responses fit a particular item response model. The person-fit index used by the authors was *lz* (Drasgow, Levine, & Williams, 1985), the standardized version of the statistic *l* (Levine & Rubin, 1979).

Next, using the response matrix of only those examinees that have been flagged as aberrant (using a cutoff established by simulations), a vector of item-level aberrance scores is created for each examinee. Item-level aberrance scores are calculated using another person-fit index, Wright and Stone's (1979) unweighted mean square:

$$U = \frac{1}{n} \sum_{j=1}^n \frac{(u_{ij} - P(u_{ij}))^2}{P(u_{ij}=1)P(u_{ij}=0)}. \quad (11)$$

Values of the statistic *U* that are greater than 1.3 indicate that an examinee's response patterns are not described well by the item response model. When examinees have preknowledge of a subset of items, it is likely that they will have a higher value of *U* than they otherwise would, characterized by correctly answering the compromised items more often than would be predicted by their estimated ability. Item-level aberrance scores are calculated using the component that is summed over all of the items, *V_{ij}*, which is calculated as:

$$V_{ij} = \frac{(u_{ij} - P(u_{ij}))^2}{P(u_{ij}=1)P(u_{ij}=0)}. \quad (12)$$

When *V_{ij}* > threshold, the aberrance score is coded as 1. When *V_{ij}* < threshold, the aberrance score is coded as 0. Thus, an *n* × *j* matrix (where *n* is the number of examinees initially flagged by *lz*) composed of 1s and 0s is used to indicate which examinees had aberrant responses to which items.

Next, a similarity measure is developed between each pair of examinees based on patterns of aberrance scores. The authors recommended standardizing item scores across items, then taking the dot product of each pair of examinees' standardized aberrance score vectors. The standardization before calculating the dot products between all pairs of examinees makes the resulting *n* × *n* similarity matrix a correlation matrix.

When a similarity matrix contains correlation coefficients between objects, factor loadings from a Q-type factor analysis can be used to cluster examinees already flagged from the initial person-fit analysis who have similar item-level aberrance scores, which could potentially be due to having preknowledge of the same set of items.

The Zhang et al. (2011) study was motivated by a known security breach and included a real data example along with a simulation. The authors explained how their simulation mirrors the real data example for variables that were known to them (i.e., test length and distribution of item difficulty estimates), and how they varied some variables that were unknown to them but could have influenced the method's performance (i.e., distribution of ability for examinees with preknowledge, number of aberrant groups, size of aberrant groups, and number of compromised items). In the real data analysis, 4.2% of over 2,000 examinees were initially flagged in the person-fit analysis, and of those, only 18% to 20% were grouped by the factor analysis.

DETECTING BOTH INDIVIDUALS AND ITEMS

Differential Person Functioning and Differential Item Functioning

O'Leary and Smith (this volume) developed a two-step approach to first identify individuals who have likely benefited from preknowledge, then used those classifications to identify the specific items likely to be compromised. To use this method, practitioners must imbed some type of "security" items into the exam, which could be any set of items known not to be compromised. Examinees can then be grouped into two groups based on the analysis: those who likely had preknowledge of some subset of operational items and those who likely did not have preknowledge. Differential Item Functioning (DIF) can then be conducted for each group to identify specific items that may have been compromised.

Smith and Davis-Becker (2011) described using different criteria for flagging individuals under different circumstances. While the authors only briefly mentioned this in their methodology, their procedure highlights some important considerations for detection of item preknowledge in general. The data used in their study come from an international certification exam program with a "serious security problem as indicated by the large number of examinees who scored very well on the exam in an extremely short amount of time" (p. 5). DPF statistics were first used in the calibration of items, where examinees with DPF contrasts greater than 1 logit were excluded from the analysis. Then, for the purpose of flagging those who may have benefitted from preknowledge, examinees with DPF contrasts greater than 3 logits were flagged. For the purposes of item calibration, a lower threshold for flagging is used that tolerates a higher false positive rate to exclude more examinees with preknowledge for purposes of item parameter estimation. However, in the context of flagging individual examinees, a higher threshold for flagging is necessary because flagging an individual examinee is a very serious undertaking. This process highlights the need to lessen the contamination resulting from examinee preknowledge in item parameter estimates before moving forward with further analyses, including flagging individual examinees.

FLOR Log Odds Ratio Index

McLeod and Schnipke (2006) introduced an approach to person and item fit that is similar in approach to O'Leary and Smith (this volume) in that their methodology first

identifies suspect examinees, then uses those classifications to identify suspect items. The first step to identify examinees is the calculation of the FLOR log odds ratio index (McLeod et al., 2003) based on estimated IRT model parameters, examinees' response data, and several possible models specifying the probability that an item has been compromised. The log odds ratio index is first calculated at the examinee level and is based on two models for item response: one that assumes the examinee is honest (\bar{s}) and one that assumes the examinee has benefited from preknowledge (s) and therefore has an inflated estimated ability. Responses in the absence of preknowledge are modeled with the 3PL model (Birnbaum, 1968):

$$p(u_i = 1|\bar{s}, \theta) = c_i + \frac{(1 - c_i)}{1 + \exp[-1.7a_i(\theta - b_i)]} \quad (13)$$

where a_i is the discrimination parameter, c_i is the guessing parameter, and b_i is the difficulty parameter.

Responses to compromised items by examinees who have preknowledge are modeled with a modified 3PL model that also models the probability each item is compromised with some function, $p(m_i)$:

$$\begin{aligned} p(u_i = 1|s, \theta) &= (1)p(m_i) + (1 - p(m_i))p(u_i = 1|\bar{s}, \theta) \\ &= p(m_i) + \left\{ c_i + \frac{(1 - c_i)}{1 + \exp[-1.7a_i(\theta - b_i)]} \right\} - p(m_i) \left\{ c_i + \frac{(1 - c_i)}{1 + \exp[-1.7a_i(\theta - b_i)]} \right\}. \end{aligned} \quad (14)$$

Using these response models, the probability that an individual has preknowledge given the response to Item 1 can be written as:

$$p(s|u_1) = \frac{\int p(u_1|s, \theta)p(s)p(\theta)d\theta}{\int p(u_1|s, \theta)p(s)p(\theta)d\theta + \int p(u_1|\bar{s}, \theta)p(\bar{s})p(\theta)d\theta}. \quad (15)$$

After the first item, it is no longer assumed that preknowledge status (s vs. \bar{s}) and θ are independent, because the distribution of θ has been updated based on examinees' responses that may reflect item preknowledge. Therefore, $p(s)$ after the first item can be written as:

$$p(s|u_1, u_2) = \frac{p(u_2|s, \theta)p(s, \theta|u_1)d\theta}{\int p(u_2|s, \theta)p(s, \theta|u_1)d\theta + \int p(u_2|\bar{s}, \theta)p(\bar{s}, \theta|u_1)d\theta}. \quad (16)$$

Thus, the posterior probability that an examinee is benefitting from preknowledge after n items is:

$$p(s|u_1, \dots, u_n) \propto \int p(u_n|s, \theta)p(s, \theta|u_1, \dots, u_{n-1})d\theta. \quad (17)$$

To compute a more interpretable index, the above quantity can be used to calculate the log odds ratio for each examinee. The statistic l_o is the base 10 logarithm of the ratio of the current odds (after observing item responses to n items) to the prior odds that the individual benefitted from item preknowledge, $p(s)$.

$$l_o = \log_{10} \left[\frac{p(s|u_1, \dots, u_n) / [1 - p(s|u_1, \dots, u_n)]}{p(s) / [1 - p(s)]} \right] \quad (18)$$

Negative values of the log odds ratio indicate that an examinee is less likely to have benefitted from item preknowledge compared to the prior probability, and positive values indicate that the examinee is more likely to have benefitted from item preknowledge compared to the prior probability. This statistic is similar to that used by Obregon (2013), but it represents the log odds of an individual examinee having preknowledge rather than the log odds that an item is compromised. By taking the base 10 logarithm of the odds ratio, we can interpret a final log odds ratio of 1 as meaning that we have 10 times more suspicion than there was before we observed the n item responses that the examinee has benefitted from item preknowledge.

In McLeod et al.'s earlier study (2003), calculation of several different FLOR indices varied according to how the probability of item compromise (i.e., $p(m_i)$) was modeled by either using a constant probability of compromise, weighting the probability by difficulty (where the most difficult items have a higher probability of being compromised), or employing an empirical model using the item selection algorithm and assuming several numbers of memorizing sources at varying true scores. McLeod and Schnipke (2006) chose to use FLOR3 in their study for identifying both suspect individuals and items, where the log odds ratio is computed using $p(m_i | b_i) = \frac{1}{(1 + \exp(-b_i))}$. Given

this calculation of $p(m_i)$, we can see that the more difficult items are assumed to have a higher prior probability of being compromised.

After computing FLOR3 values for every examinee, the FLOR_i statistic was computed at the item level to flag which items may have been compromised. Examinees were classified into two groups—a suspect group and a nonsuspect group—based on their FLOR3 values. Examinees with FLOR3 values less than zero were classified in the nonsuspect group, and examinees with FLOR3 values greater than zero were classified in the suspect group. The FLOR_i index is calculated as:

$$\text{FLOR}_i = \log_{10} \left[\frac{p(u=1|s) / [1 - p(u=1|s)]}{p(u=1|n) / [1 - p(u=1|n)]} \right] \quad (19)$$

where $p(u=1|s)$ is the proportion of examinees in the suspect group that answered the item correctly, and $p(u=1|n)$ is the proportion of examinees in the nonsuspect group that answered the item correctly. If the proportion of examinees answering the item correctly is the same for the suspect group and the nonsuspect group, then the FLOR_i index will equal zero. Values greater than zero indicate that the proportion of examinees in the suspect group are answering the item correctly more often than examinees in the nonsuspect group.

The authors simulated various types and amounts of examinee preknowledge by varying the type of source (i.e., “professional” vs. “regular,” where professional sources are of higher ability and try to memorize the most difficult items, and regular sources are of average ability), number of sources (i.e., 2 vs. 6), and percentage of examinees that are beneficiaries of the stolen content (i.e., 10% vs. 25%). They also included a null condition in which no examinees are beneficiaries of stolen content. Response data were generated for a population of examinees ranging in ability from -3 to 3 in increments of 0.5, where response data for 1,000 examinees were simulated at each

increment, for a total of 13,000 examinees per replication. There were a total of 10 replications.

The authors included a table that broke down the number of flagged examinees by condition and true theta classification. The table showed that examinees with higher true abilities were overwhelmingly more likely to be flagged in comparison to examinees with lower true abilities. For example, in the null condition where none of the examinees have had preknowledge of any of the items, 0.39% of examinees who had a true theta of -3.0 were incorrectly flagged as having preknowledge, whereas 89.5% of examinees who had a true theta of 3.0 were incorrectly flagged as having preknowledge. This result is very disconcerting, especially when we consider how this information is used to then calculate FLOR_i . The input values for the FLOR_i statistic include the proportion of the suspect group that answered the item correctly and the proportion of the nonsuspect group that answered the item correctly. If the ability distribution of the suspect group is always higher than the ability distribution of the nonsuspect group, we would expect FLOR_i to always be positive. And, all else equal, we would expect FLOR_i to have a greater positive magnitude for a difficult item than an easy item.

Model performance for results of individual flagging were illustrated best in the 2003 study, whereas results for item flagging are illustrated best in the 2006 study. Receiver operating characteristic curves (ROCs) showing false positive detection rates and corresponding false negative detection for individual examinees using this method indicated that modeling the probability of item compromise according to item difficulty and the empirical model works best. Of the simulated conditions for identifying compromised items, ROCs showed that increasing the number of sources (the highest number simulated was six sources) and increasing the number of beneficiaries (the highest percentage of beneficiaries simulated was 25%) resulted in the best model performance.

Modeling Speed and Accuracy on Test Items

When examinees have preknowledge of live items, they are likely to answer those items more quickly than they would have without preknowledge because they eliminate the cognitive processes required to respond to the items—they just need to recall the correct answers. Several researchers have used response time modeling in addition to item response modeling to detect both examinees that may have benefitted from preknowledge and compromised items. Because many other detection methods in the area of item preknowledge may flag examinees or items for aberrances that do not result from item preknowledge, using methods that analyze response time in addition to item response may be beneficial. Response time analysis can focus on detecting aberrant item performance or aberrant examinee performance, where an item may be flagged if many examinees have unexpectedly short response times on that item and an individual examinee may be flagged if he or she has unexpectedly short response times on many items. Boughton, Smith, and Ren (this volume) provide a detailed description of methods utilizing response times applied to the common licensure dataset.

DISCUSSION

Security threats in the form of item compromise and examinee preknowledge of live exam content have the potential to undermine the mission of an exam program and call into question the validity of its scores, inferences, and/or credentials issued.

Detection methods have been developed in response to these emerging threats to identify potentially compromised items and individuals who may have benefitted from preknowledge, but detection efforts face a variety of challenges in practice. For example, response patterns consistent with the hypothesis of preknowledge could have other explanations besides preknowledge, and variables that influence a method's performance may not be known to the practitioner (e.g., percent of compromised items, percent of examinees who have benefitted from preknowledge).

The literature reviewed here presents several broad themes relating to the effective use of statistical methods to identify compromised items and/or examinees who have benefitted from preknowledge, and different methods will be useful in different contexts. I have organized the discussion of methods to highlight the circumstances that could lead to the utilization of a particular method, and to draw similarities and differences between the various methods.

Tying this body of research together, there are several rules of thumb we can follow for future research and evaluating methods for use in a practical context:

*When Analyzing Real Data, First Conduct Simulation Studies
That Mimic the Characteristics of the Real Data as Much as Possible*

Zhang et al. (2011) illustrate how this strategy can be carried out in practice. Because many factors can influence performance of methods to detect aberrances resulting from item compromise and examinee preknowledge, this strategy can help practitioners interpret results from real data. Han and Hambleton (2004) also employ this strategy.

*Efforts Should be Made to Remove Contamination Due to Item
Preknowledge from Parameter Estimates Used in a Method*

Several studies discussed in this chapter include some attempt to minimize the contamination of parameter estimates due to item preknowledge. Eckerly et al. (2015) address this most explicitly, while Belov (2013) and Smith and Davis-Becker (2011) briefly mention this in their methodology. Wollack & Maynes (this volume) also show through simulations how contamination affects results. Future research should address how this goal can better be accomplished, keeping in mind that contamination is most severe with high amounts of compromise.

*It Is Important to Know How a Method Performs
When No Examinees Have Preknowledge*

Most simulation studies on methods to detect item compromise and examinee preknowledge currently only show how a method performs in the presence of some simulated item preknowledge, which is of course the main purpose of these types of studies. However, every exam program wants to find itself in the scenario where there actually is no item compromise or examinee preknowledge. Therefore, ensuring that methods to identify aberrances of these types do not perform in unanticipated ways when there is actually no preknowledge is imperative. Eckerly et al. (2015) and Han and Hambleton (2004) illustrate ways this can be done with both the scale-purified DGM and moving averages of p -values.

*When Conducting Simulation Studies, Simulate Conditions
That Highlight How the Model Performs in Realistic Scenarios*

Simulation studies should not only focus on conditions where the model is likely to perform optimally but also on conditions that represent realistic scenarios that exam programs are likely to experience. Simulations should present rigorous tests by evaluating a method's efficacy in conditions likely to present worst case (but plausible) scenarios. If a method can perform adequately in the least favorable conditions, we can safely assume it will perform better in practice. Eckerly and Wollack (2013) show how changes in assumptions about the ability distribution of examinees with preknowledge and plausible scenarios of user input misspecification dramatically change model performance. Also, some studies using simulations make the assumption that difficult exam content is more likely to be compromised, even though there has not been any research to indicate that this is an accurate assumption. If that questionable assumption used to simulate which items have been compromised is then used to model the probability that an item is compromised, as was done by McLeod et al. (2003) and McLeod and Schnipke (2006), the method will be set up for optimal performance that may not be realistic. Similarly, Obregon (2013) makes an assumption in his simulation that lower ability candidates are more likely to have preknowledge, then models the prior probability that an examinee has preknowledge in the exact same manner that preknowledge was simulated. Conducting simulations with these types of assumptions is helpful for showing model performance when the assumptions are in fact correct, but until these assumptions are shown to be valid, we should be careful about using them to infer overly optimistic performance.

Not All Flags Are Created Equal

For any given method, a flag in one situation may spur investigation or corrective actions on the part of an exam program, where another flag may be deemed an errant flag and is not a cause for concern. Understanding the method used well enough to make an informed decision is imperative. For example, Eckerly et al. (2015) compare flag rates before and after a known breach to determine if flags after the breach are likely to be errant. Nonstatistical information can also be used to complement analysis of flags.

Flagging Criteria Are Not “One Size Fits All”

The goal of any investigation should be taken into account when establishing flagging criteria. If the goal is to cast a wide net to gain an understanding about how widespread a preknowledge problem might be, a lower threshold that leads to more false positives may be acceptable. If flagging criteria need to be established for multiple steps (e.g., an intermediate step for purposes of scale-purification as was shown in Eckerly et al. (2015) and Smith and Davis-Becker (2011), as well as a final step for identifying individuals with preknowledge), there may be different considerations for each step. Flagging criteria that do not lend themselves to easily understood probability statements (e.g., flagging criteria for the log odds ratio statistics) may need to be customized to specific exam programs based on empirical distributions of statistics from a large number of past data sets. Additionally, statistics that are used along with collateral information (e.g., information from a tip line) might require lower thresholds than statistics that are used in isolation.

CONCLUSION

Keeping these considerations in mind, I offer one final recommendation: practitioners should not wait until they think they have a security problem to decide how to manage it. In fact, if an exam program has reason to believe that preventative measures are currently keeping item content secure, it is the perfect time to start investigating how different methods perform for that specific program in practice. For example, if one is using the log odds ratio statistic (Obregon, 2013) to flag items that may have been compromised, it could be valuable to look at the distribution of statistics over time at the exam level to establish a baseline as to what can be expected under conditions of no compromise. Simulations can be helpful to establish this baseline, but even the best efforts to mimic real response data will fall short. Establishing appropriate flagging criteria can be greatly aided by these types of analyses.

While the area of research dedicated to detecting item compromise and examinee preknowledge is fairly new, there have been many promising developments in recent years. Because it is not likely that these threats will disappear in the near future, further research to develop new methods and refine existing ones will greatly benefit the testing community. Looking forward, researchers can build on the current body of research with additional goals to develop methods that have more acceptable and consistent Type I error control, minimize or remove contamination of parameter estimates due to item compromise and examinee preknowledge, and better lend themselves to practical use as part of an exam program's comprehensive plan to address exam security threats.

REFERENCES

- Belov, D. (2012). *Detection of large-scale item preknowledge in computerized adaptive testing via Kullback–Leibler divergence*. LSAC Research Report (RR 12-01).
- Belov, D. (2013). Detection of test collusion via Kullback-Leibler divergence. *Journal of Educational Measurement*, 50, 141–163.
- Belov, D. (this volume). Identification of item preknowledge by the methods of information theory and combinatorial optimization. In G. J. Cizek and J. A. Wollack, Eds., *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Eds., *Statistical theories of mental test scores* (chaps. 17–20). Reading, MA: Addison Wesley.
- Boughton, K. A., Smith, J., & Ren, H. (this volume). Using response time data to detect compromised items and/or people. In G. J. Cizek and J. A. Wollack, Eds., *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polytomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67–86.
- Eckerly, C. A., Babcock, B., & Wollack, J. A. (2015). *Preknowledge detection using a scale-purified deterministic gated IRT model*. Paper presented at the annual meeting of the National Conference on Measurement in Education, Chicago, IL.
- Eckerly, C. A., & Wollack, J. A. (2013, October). *Detecting examinees with preknowledge: Examining the robustness of the deterministic gated item response theory model*. Second Annual Conference on Statistical Detection of Potential Test Fraud, Madison, WI.
- Fitzgerald, C. T., & Mulkey, J. R. (2013). Security planning, training, and monitoring. In J. A. Wollack and J. J. Fremer, Eds., *Handbook of test security* (pp. 127–146). New York, NY: Routledge.
- Foster, D. (2013). Security issues in technology-based testing. In J. A. Wollack and J. J. Fremer, Eds., *Handbook of test security* (pp. 39–83). New York, NY: Routledge.
- Han, N. (2003). *Using moving averages to assess test and item security in computer based testing* (Research Report No. 468). Amherst, MA: University of Massachusetts, School of Education, Center for Educational Assessment.

- Han, N., & Hambleton, R. (2004). *Detecting exposed test items in computer-based testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Hornby, L. (2011, July 27). *Gaming the GRE test in China, with a little online help*. Retrieved October 18, 2015, from www.reuters.com/article/2011/07/27/us-china-testing-cheating-idUSTRE76Q19R20110727.
- Kyle, T. (2002, August 9). *Cheating scandal rocks GRE, ETS*. Retrieved October 18, 2015, from <http://thedartmouth.com/2002/08/09/cheating-scandal-rocks-gre-ets/>.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics*, 4, 269–290.
- Maynes, D. (2009). *Caveon speaks out on IT exam security: The last five years*. Retrieved October 18, 2015 from www.caveon.com/articles/IT_Exam_Security.pdf.
- McLeod, L., & Lewis, C. (1999). Detecting item memorization in CAT environment. *Applied Psychological Measurement*, 23, 147–159.
- McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, 27, 121–137.
- McLeod, L. D., & Schnipke, D. L. (2006). *Detecting items that have been memorized*. Newtown, PA: Law School Admission Council.
- Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, 8, 261–272.
- Obregon, P. (2013). *A Bayesian approach to detecting compromised items*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- O'Leary, L. S., & Smith, R. W. (this volume). Detecting candidate preknowledge and compromised content using differential person and item functioning. In G. J. Cizek and J. A. Wollack, Eds., *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.
- Scicchitano, A. R., & Meade, R. D. (2013). Physical security at test centers and the testing company. In J. A. Wollack and J. J. Fremer, Eds., *Handbook of test security* (pp. 147–179). New York, NY: Routledge.
- Shu, Z., Leucht, R., & Henson, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika*, 78, 481–497.
- Skorupski, W. P., & Wainer, H. (this volume). The case for Bayesian methods when investigating test fraud. In G. J. Cizek and J. A. Wollack, Eds., *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.
- Smith, R. W. & Davis-Becker, S. (2011, April). *Detecting suspect candidates: An application of differential person functioning analysis*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden and C. A. W. Glas, Eds., *New developments in computerized adaptive testing: Theory and practice* (pp. 201–219). Boston: Kluwer-Nijhoff Publishing.
- Wollack, J. A. & Maynes, D. D. (this volume). Detection of test collusion using cluster analysis. In G. J. Cizek and J. A. Wollack, Eds., *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.
- Zamost, S., Griffin, D., & Ansari, A. (2012, January 13). *Exclusive: Doctors cheated on exams*. Retrieved from www.cnn.com/2012/01/13/health/prescription-for-cheating/.
- Zara, A. (2006). *Defining item compromise*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Zhang, Y., Searcy, C. A., & Horn, L. (2011). *Mapping clusters of aberrant patterns in item responses*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

6

DETECTION OF TEST COLLUSION USING CLUSTER ANALYSIS

James A. Wollack and Dennis D. Maynes

Over the past 10 to 15 years, the educational measurement literature has seen a marked increase in the number of articles related to test security. In particular, substantial attention has been focused on detection of answer copying (Belov & Armstrong, 2009; Sotaridona & Meijer, 2003; van der Linden & Sotaridona, 2004, 2006; Wesolowsky, 2000; Wollack, 1997), prevention of item disclosure in a computerized adaptive testing environment (Chang & Ying, 1999; Chen & Lei, 2005; Davey & Parshall, 1995; Doong, 2009; Impara, Kingsbury, Maynes, & Fitzgerald, 2005; Leung, Chang, & Hau, 2002; Nering, Davey, & Thompson, 1998; Stocking & Lewis, 1998; Stocking & Swanson, 1993; van der Linden, 2003; Yi & Chang, 2003), and detection of item preknowledge, typically using response-time data (Belov, 2014, 2015; McLeod & Lewis, 1999; Meijer & Sotaridona, 2006; Shu, Henson, & Leucht, 2013; van der Linden, 2009; van der Linden & Guo, 2008; van der Linden & van Krimpen-Stoop, 2003; and van Krimpen-Stoop & Meijer, 2001). One area that has been underrepresented in the test security literature is the detection of test collusion among multiple examinees, or sets of examinees with unusual answer patterns in common. The lack of research in this area is somewhat surprising in light of the fact that this type of organized test security breach can seriously jeopardize the integrity of testing programs. Examples of collusion include illicit coaching by a teacher or test-prep school, examinees accessing stolen test content posted on the World Wide Web, examinees communicating about test answers during an exam, examinees harvesting and sharing exam content using e-mail or the Internet, and teachers or administrators changing answers after tests have been administered.¹

To date, the only published method to detect collusion from item response patterns was developed by Jacob and Levitt (JL, 2003). The JL method specifically focused on teacher cheating but could conceivably extend to other groups of examinees for which collusion was possible. The JL method identifies collusion through consideration of two summary scores, one relating to the overall atypicality of answer strings within a classroom and the other relating to unexpected score fluctuations. To derive the answer strings summary measure, within a particular year and subject area, individual

classrooms were ranked separately with respect to (a) the likelihood of the most unusual block of identical answers given by a set of students (taking into consideration both class size and test length), (b) a measure of within-classroom correlation across all items, (c) the variance of within-question correlations across all test items, and (d) a comparison of the students' answers to those provided by students in different classes who took the same tests and had the same raw score. The overall answer strings summary measure was taken as the sum of squared ranks for each of these four indexes.

The unexpected score fluctuations summary measure was computed in similar fashion. Within a particular year and subject area, individual classrooms were separately ranked on the gain in percentile rank (from the previous year), for two consecutive years (i.e., $gain1 = \text{rank}(pctile_{c,b,t_0+1} - pctile_{c,b,t_0})$ and $gain2 = \text{rank}(pctile_{c,b,t_0+2} - pctile_{c,b,t_0+1})$), where c denotes classroom, b denotes subject, and t_0 denotes the baseline year. The composite is taken as the sum of squares for $gain1$ and $(1 - gain2)$.

The JL method detects collusion by considering the joint distribution of these two summary statistics. Specifically, for each classroom, the percentile rank of the answer strings summary index is plotted against the probability that the classroom exceeded the 95th percentile with respect to the measure of unusual score fluctuations. Classrooms with high values on both summary measures are flagged as potential instances of cheating.

The JL method appears to be a promising approach for identifying collusion; however, it does have a few drawbacks that limit its potential utility. First, because JL is a within-class design, it requires the specification of which students are in which class. For purposes of teacher cheating (which, in fairness, is what the JL study was intended to address), it is not difficult to identify which students are in which classrooms. However, as a general approach to collusion detection, it is not desirable because several types of collusion (e.g., obtaining stolen test content through a website or test-prep company) have the potential to affect unknown groups of individuals at multiple test sites. A second drawback is that the JL method does not identify individual examinees involved in the collusion. Again, if teacher cheating is what is suspected, examinees are innocent victims, so the identification of examinees is not necessary. However, for many other types of collusion, it is important to identify the examinees involved so as to (a) facilitate a thorough investigation into whether or not collusion occurred, and (b) take appropriate action against those individuals, should it be deemed necessary. Finally, by virtue of analyzing the gain scores, the JL method requires complete data over a three-year (or three testing window) period. With the exception of school accountability testing programs which require repeatedly testing students over a multiple-year period, this constraint would result in excluding data for many examinees. Furthermore, for most programs, particularly licensure and certification programs, repeat testers do not constitute a representative sample, because those who passed or scored high are significantly less likely to retest. As a result, even if it were possible to identify a large enough sample of examinees who tested over three consecutive administrations, the sample would not include any examinees who passed the exam. A gain score approach, therefore, focuses only on unsuccessful cheaters, because those who work together, as a group, and successfully pass the exam will not be retesting.

This chapter aims to fill a void in the literature by presenting a new approach to detect clusters of examinees engaged in test collusion. The approach described herein is applicable to any type of collusion, including teacher cheating, test coaching (either by a classroom teacher or from a review course), systematic answer sharing during the test (e.g., multiple-examinee copying or communication systems involving multiple

examinees), and use of harvested items. Moreover, this method does not require that the groups of potentially contaminated examinees be identified *a priori*, can be applied to data from a single test administration, and possesses the ability to identify individuals whose test scores are of questionable validity.

The organization of this chapter is as follows: First, the new method for detecting test collusion will be presented. Next, a new model for simulating varying magnitudes of collusion will be presented, followed by a simulation study used to investigate the statistical properties of the collusion model under realistic conditions, where the dataset includes varying amounts of both honest and contaminated individuals. Finally, the model will be applied to the common licensure dataset to identify groups of examinees who may have engaged in collusion.

DETECTING COLLUSION

When examinees work together on tests, are illegally coached on certain test material, have access to live exam materials, or have their answers systematically changed afterwards, it is expected that their item response strings will be more similar to each other's than they would otherwise have been. Two different categories of statistical indexes exist for identifying pairs of examinees with unusually similar tests: answer copying indexes and answer similarity indexes. Answer copying indexes (e.g., Belov & Armstrong, 2009; Frary, Tideman, & Watts, 1977; Sotaridona & Meijer, 2003; van der Linden & Sotaridona, 2004; Wollack, 1997) postulate a directional hypothesis that one examinee copied answers from a particular second examinee. Answer copying index values for examinee A copying answers from B are different than (though not independent of) index values investigating B copying from A. In contrast, answer similarity indexes (Maynes, 2014; van der Linden & Sotaridona, 2006; Wesolowsky, 2000) provide an estimate of the likelihood of two examinees' shared responses, without specification of one examinee as an alleged copier and the other as a potential source.

In instances where statistical measures are used to corroborate a specific allegation that one examinee copied answers from another (as might be the case if a proctor observed the cheating during the exam), answer copying indexes are preferable both on theoretical grounds and because they are generally more powerful. However, in exploratory research, similarity indexes are preferable because the goal is often to trigger an investigation.

For purposes of detecting collusion, answer similarity indexes are more useful. First, collusion is a much broader category of test compromise than is answer copying. Whereas answer copying requires that one or more examinees serve as source, with collusion, it is possible that the true source is not even part of the dataset. For example, an instructor who collects information from students after they test, then uses that information to coach future test takers (as was discovered a few years ago to be common practice within radiology residency programs; Zamost, Griffin, & Ansari, 2012) will create unusual homogeneity of test responses among his or her students. Yet, the instructor does not produce a source item response vector against which to compare those students' answers.

The second, and more important reason to prefer answer similarity indexes, however, is that they are symmetric (i.e., the index value between examinees A and B is identical to the value between examinees B and A). Consequently, the similarity between all possible pairs can be determined, and this triangular matrix can be subjected to a variety of clustering methods for purposes of identifying sets of interrelated examinees.

In this paper, we discuss the use of the nearest neighbor (or single linkage) clustering method. In using the nearest neighbor clustering method with similarity data, clusters S and T , respectively containing examinees s_i and t_j ($i = 1, \dots, n_1, j = 1, \dots, n_2$),

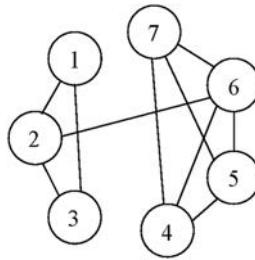


Figure 6.1 Illustration of Nearest Neighbor Clustering Approach

are clustered together if the similarity index for pair $[s_i, t_j]$, $S(U_{si}, U_{tj})$, exceeds some prespecified threshold, δ , for at least one $[s_i, t_j]$. This clustering approach is shown most clearly in Figure 6.1, where solid lines indicate examinee pairs for whom $S(U_{si}, U_{tj})$ exceeds δ (the clustering threshold). Clearly, examinees 1, 2, and 3 are strongly related; similarly, examinees 4, 5, 6, and 7 are strongly related. Relationships between members of these two sets appear weaker; however, because $S(U_2, U_6)$ is also significant, all seven examinees will be clustered together.

Implementation of the nearest neighbor clustering method requires two elements: selection of a statistical tool to estimate $S(U_{si}, U_{tj})$ and specification of the clustering threshold δ . Although any of several similarity indexes could have been used (e.g., van der Linden & Sotaridona, 2006; Wesolowsky, 2000), for purposes of this study, we computed the $M4$ answer similarity index (Maynes, 2014; this volume) between all possible pairs of examinees and took those individual values as our measures of $S(U_{si}, U_{tj})$. Because $M4$ index values are reported on a log base-ten metric (i.e., $M4 = -\log_{10}(p)$), for this study, we set $\delta = 1.301$, meaning that a particular pair of examinees was considered sufficiently similar if the $M4$ index was extreme enough to be statistically significant at the $\alpha = .05$ level.

Computation of M4

Van der Linden and Sotaridona (2006) developed an answer similarity statistic that used the family of generalized binomial distributions to derive the exact null distribution of the number of matching answers between a pair of independently working examinees, under the assumption that responses for honest examinees fit an item response model such as the nominal response model (NRM; Bock, 1972). Whereas van der Linden and Sotaridona's generalized binomial model considered only the number of items with matching choices and the number for which the choices did not match, the $M4$ statistic (Maynes, 2014; this volume) decomposes the number of matching answers into its constituent parts: the number of identical correct answers and the number of identical incorrect answers. Therefore, $M4$ uses a generalized trinomial distribution to derive the exact distribution of the number of identically correct and incorrect answers.

We will assume (as van der Linden and Sotaridona did) that for an independently working examinee j , the response probabilities can be described by the NRM (Bock, 1972), such that

$$\pi_{ji_a} = \frac{\exp(\zeta_{i_a} + \lambda_{i_a} \theta_j)}{\sum_{k=1}^A \exp(\zeta_{i_k} + \lambda_{i_k} \theta_j)}, \quad (1)$$

where π_{jia} is the probability of examinee j selecting response a to item i , ζ_{ia} and λ_{ia} are the nominal response model intercept and slope parameters corresponding to alternative a of item i , θ_j is the ability level of examinee j , and A is the total number of response options for item i .

Under the assumption that examinees j and s are working independently of each other, the joint probability of these examinees' responses to item i , π_{jsi} , is given by the product of the probabilities of j selecting a and s selecting a' , as shown below:

$$\pi(r_{ji} = a, r_{si} = a' | \theta_j, \theta_s) = \pi_{jsi} = \pi_{jia} \pi_{si_a'}, \quad (2)$$

where r_{ji} is the response to item i provided by examinee j and r_{si} is the response provided by examinee s . Therefore, we will define the probability of examinees j and s jointly selecting the correct answer to item i as

$$P_i = \hat{\pi}_{jia} \hat{\pi}_{si_a} I(a = r_k), \quad (3)$$

where $\hat{\pi}_{jia}$ is the empirical estimate of the probability in Equation (1), r_k denotes the keyed alternative for item i , and $I(\cdot)$ is an indicator function that equals 1 if the statement in parentheses is true, and 0 otherwise.

Similarly, we define

$$Q_i = \sum_{a=1}^A \hat{\pi}_{jia} \hat{\pi}_{si_a} I(a \neq r_k) \quad (4)$$

as the probability of examinees j and s selecting the same incorrect alternative for item i , and

$$R_i = 1 - P_i - Q_i = \sum_{a=1}^A \sum_{a'=1}^A \hat{\pi}_{jia} \hat{\pi}_{si_a'} I(a \neq a') \quad (5)$$

as the probability of examinees j and s jointly selecting different alternatives for item i .

Let $f_t(m, n)$ be the probability of m matching correct and n matching incorrect answers on t items. Therefore, for a test with $t = 2$ items, the possible outcomes for (m, n) and their corresponding probabilities are given in Table 6.1. For tests with more than two items, the probabilities of various outcomes can be found using the following recursion formula:

$$f_t(m, n) = P_t f_{t-1}(m-1, n) + Q_t f_{t-1}(m, n-1) + R_t f_{t-1}(m, n), \quad (6)$$

subject to the boundary condition that $f_0(0, 0) = 1$ when $(m = n = 0)$ and $f_0(m, n) = 0$, otherwise for all values of m and n .

Table 6.1 Explication of the Trinomial Probabilities for a Two-Item Test

(m, n)	Probability of Outcome
(0, 0)	$R_1 R_2$
(0, 1)	$Q_1 R_2 + R_1 Q_2$
(1, 0)	$P_1 R_2 + R_1 P_2$
(1, 1)	$P_1 Q_2 + Q_1 P_2$
(2, 0)	$P_1 P_2$
(0, 2)	$Q_1 Q_2$

Note: (m, n) is the number of matching correct answers and the number of identically incorrect answers between a pair of examinees.

A Note on the Clustering Threshold δ

As mentioned previously, for this study, δ was set at 1.301 so that only those $M4$ indexes that were statistically significant at the $\alpha = .05$ level were flagged. However, when multiple similarity indexes are computed, a multiple comparisons procedure needs to be used to correct for the increased likelihood of committing a Type I error (Maynes, 2013, this volume; Wesolowsky, 2000; Wollack, Cohen, & Serlin, 2001). In this instance, we elected to control the Type I error rate at the examinee level,² that is, so that the probability of falsely flagging an examinee for possible cheating was equal to α . In computing all possible pairs, each pair is involved in $(N - 1)$ indexes. However, a multiple comparisons correction that tests each index for significance using $\alpha/(N - 1)$ will result in a test that is too conservative, because each time one of these indexes is significant, two examinees are flagged, not just one. Therefore, the Type I error contribution towards any one individual needs to be cut in half. Therefore, the appropriate correction is $(N - 1)/2$. To implement this correction, prior to taking the negative log of the p -value (recall that $M4 = -\log_{10}(p)$), we evaluated the upper tail for the maximum order statistic from a uniform distribution using $(N - 1)/2$ as the sample size, which is almost identical to the Bonferroni correction when p is very small (Maynes, 2009). Hence, for purposes of this chapter, $M4 = -\log_{10}(1 - (1 - p)^{(N - 1)/2})$. In this way, $M4$ values greater than 1.301 reflect an $\alpha = .05$ criterion, after correcting for multiple comparisons at the examinee level.

SIMULATING COLLUSION

Equation (2) provides the joint probability of (a, a') under the assumption of independence. When examinees work together, this assumption of independence is violated because the probability of the examinees matching answers will be increased. Therefore, this type of dependence can be simulated between a pair of examinees by first generating the joint probability distribution for a particular item and then inflating the probabilities of the responses matching, while deflating the probabilities of their selecting different responses. More specifically, we enhance the definition of π_{jsi} to include a collusion parameter, ϕ , as follows:

$$\pi_{jsi} = \frac{\pi_{j_a} \pi_{s_{i_a'}} \exp(\phi \cdot I(a = a'))}{\sum_{l=1}^A \sum_{m=1}^A \pi_{j_l} \pi_{s_{i_m}} \exp(\phi \cdot I(l = m))} = \frac{\exp(\zeta_{i_a} + \lambda_{i_a} \theta_j) \exp(\zeta_{i_{a'}} + \lambda_{i_{a'}} \theta_s) \exp(\phi \cdot I(a = a'))}{\sum_{l=1}^A \sum_{m=1}^A \exp(\zeta_{i_l} + \lambda_{i_l} \theta_j) \exp(\zeta_{i_m} + \lambda_{i_m} \theta_s) \exp(\phi \cdot I(l = m))}. \quad (7)$$

Under this model, for $\phi > 0$, the probability of examinees j and s selecting the same alternative for item i is inflated because the numerator is multiplied by $\exp(\phi)$ when $a = a'$. This serves to increase the probabilities associated with j and s selecting the same item alternatives. When $a \neq a'$, although the collusion term drops out of the numerator, the π_{jsi} are lower than they would be in Equation (2) because the denominator, which equals the sum of all $A \times A$ numerators, still contains the collusion term for all the $a = a'$ elements. When $\phi = 0$, this equation reduces to Equation (2).

To gain an appreciation for the way ϕ impacts π_{jsi} , Figure 6.2 shows the probability of j ($-3 \leq \theta_j \leq 3$) and s ($-3 \leq \theta_s \leq 3$) matching answers for $0.0 \leq \phi \leq 3.0$ (in increments of 0.5), averaged over parameters from a 50-item test calibrated with the nominal response model. Note that the surface for $\phi = 0.0$ shows the probabilities of answer matching when j and s work independently. Note also that the surface for $\phi = 0.5$ has been made transparent so as to facilitate seeing the

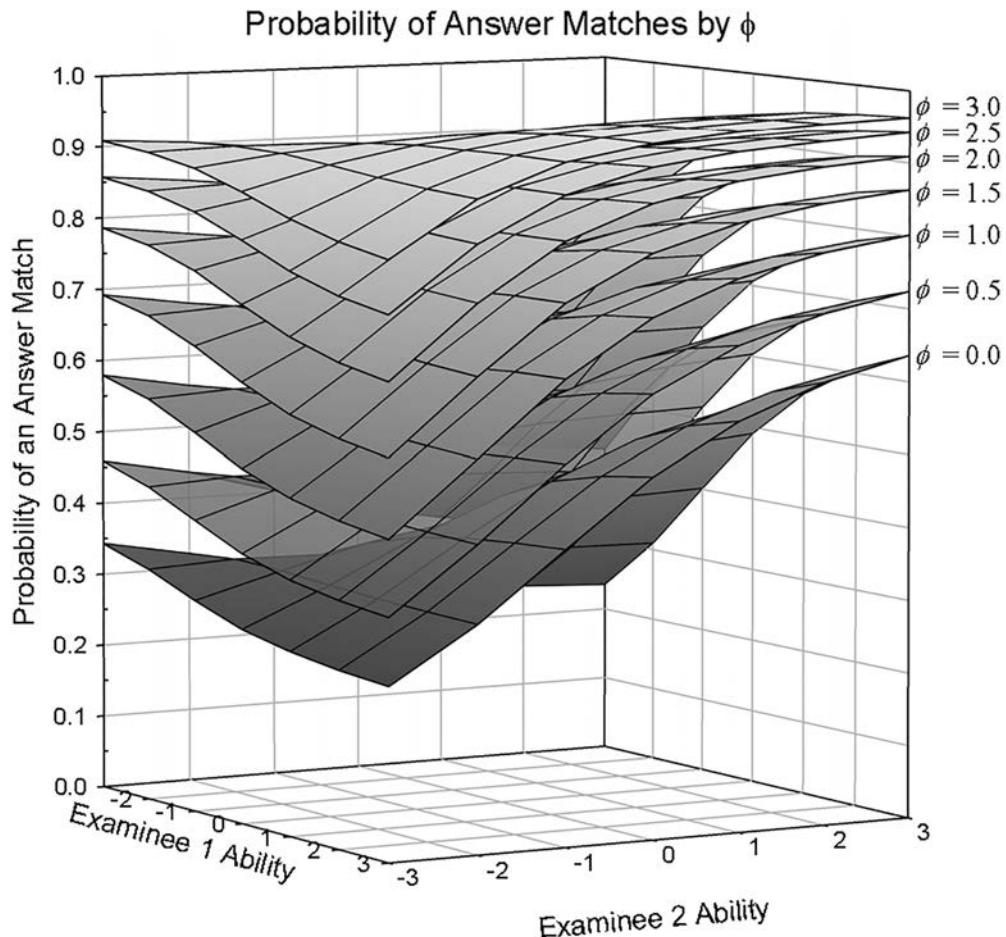


Figure 6.2 Probability of Answer Matches by ϕ

curvature in the surface for $\phi = 0$. The surfaces are symmetric around the line $\theta_j = \theta_s$, and all assume the same general shape, though they do become significantly flatter as ϕ increases.

To simulate collusion for a pair of examinees, one needs to generate two independent θ s, select ϕ , the item parameters, and the set of items on which there was collusion, and use Equation (7) to produce the $A \times A$ table of joint probabilities for each item. Once those probabilities are determined, one can derive the joint cumulative distribution, $F(a, a')$ and simulate responses by generating a random variable, w , from a $U(0, 1)$ distribution, and finding the cell in which $[F(a, a') - w]$ is minimized, while still being greater than zero.

Simulating collusion for a group of examinees, however, must proceed a little differently. The premise in group-level collusion is that all examinees in the group will be similar because they share a common source of information. If a particular class of students gets coached on test content, everyone in the class will have received the same treatment effect (i.e., the same items will be compromised, and the magnitude of compromise will be the same). However, the nature of compromise in a different classroom

with a different teacher may be quite different. Therefore, to simulate group-level collusion, for each group, g , we proceed as follows:

1. Select or randomly sample ϕ_g , item parameters, and the m items to be compromised.
2. Independently sample the mean ability (Θ_g) from a $N\left(0, \frac{1}{G}\right)$, where G is group size. Using Equation (1), for each item, compute π_{si} , the probability of the “source” selecting each item choice. Note that π_{si} will remain constant across all examinees in group g .
3. For each examinee j within g , generate an ability parameter, $\theta_{j(g)}$ from a $N(\Theta_g, 1)$ distribution. Using Equation (1), for each item, compute π_{ji} , the probability of examinee j selecting each item choice.
4. Using π_{ji} , π_{si} , ϕ_g and Equation (7), compute π_{jsi} to produce the $A \times A$ table of joint probabilities between j and s .
5. Using Equation (1) and $\theta_s = \Theta_g$, generate a target item response vector for the group, $U_g = (u_{g1}, \dots, u_{gp}, \dots, u_{gn})$. This vector defines the source information that is common to all examinees in g .
6. Using U_g , select the set of conditional probabilities in the $A \times A$ joint probability table for which $U_{si} = u_{gi}$, the target answer selected for item i .
7. Convert the conditional probabilities to cumulative probabilities, $F(u_{ji}|u_{gi})$.
8. Generate a random variable, w , from a $U(0,1)$ distribution and simulate response u_{ij} if $F(u_{j-1,i}|u_{gi}) \leq w \leq F(u_{ji})$.
9. Repeat steps 6–8 for each item.

SIMULATION STUDY

A simulation study was conducted for purposes of studying the Type I error rate and power, as well as the ability of the collusion detection model to recover known group memberships (collusion vs. independent), under realistic circumstances.

Although it is often the case that detection is improved when forensic models utilize item parameter estimates based on uncontaminated data (e.g., Wollack & Cohen, 1998), in practice, it is quite unlikely that we will encounter a large group of examinees known to have worked independently from whom we may estimate item parameters. Furthermore, in practice, if we knew prior to conducting the analysis who was part of a collusion group and who was not, the analysis itself would be unnecessary. Instead, what is likely to occur is that all examinees testing during a particular window will be analyzed together for potential collusion. Therefore, this study used models that were derived under conditions where the data were contaminated by collusion.³

From the perspective of a practitioner, combining examinees who engaged in collusion with others who did not will have several important implications. First, the item parameter estimates will be contaminated, thereby affecting the probabilities of answer matches and somewhat altering the distribution of the M_4 statistic. Second, the multiple comparisons adjustment corrects the critical values proportionate to the number of examinees being analyzed. This brings up two issues. Most obviously, with more examinees being analyzed, the uncorrected index value will need to be higher than it would need to be with fewer examinees, thereby reducing power. Also, because some subset of the examinees will belong to collusion groups, it is not possible to commit a Type I error by identifying them. However, because we do not know how many true positives are in the dataset, these examinees must also be reflected in the multiple comparisons correction. Therefore, this correction with mixed data will likely cause

the method to become conservative. Finally, combining independent examinees with collusion groups has the potential to degrade the interpretability of the clusters, since attempts to interpret Type I error cases would waste resources and potentially obscure true patterns. We will use the term *cluster integrity* to refer to the extent to which the identified clusters were easily and correctly interpreted.

Simulation Design

In this study, we simulated responses to a 50-item test, using item parameter values based on an actual test calibrated with the nominal response model. Three different independent variables were manipulated: group size ($G = 5, 10, 20$), the number of contaminated items ($m = 25, 35, 50$), and the contamination effect ($\phi = 2.0, 2.5, 3.0$). The three independent variables were fully crossed, producing 27 conditions. For each condition, we simulated responses for 100 independent collusion groups, using the process described earlier. Note that this process allows for the groups to differ with respect to overall ability level, and with respect to the specific subset of m items on which there is collusion. Because the number of contaminated individuals was affected by group size, the samples of contaminated individuals varied from 500 (when $G = 5$) to 2,000 (when $G = 20$). In addition, we simulated item responses from 5,000 independent, honest examinees (using θ_j values drawn from a $N(0, 1)$ distribution). This same sample of 5,000 honest examinees was paired with each of the 27 collusion samples to produce the datasets to be analyzed. Consequently, the sample sizes of the entire group of examinees being analyzed varied from 5,500 (for $G = 5$) to 7,000 (for $G = 20$), and the percentage of contaminated individuals varied from 9.1% to 28.6%.

Outcome Measures

Within each condition, we investigated the Type I error rate of the procedure by running the collusion model on the 5,000 independent examinees. Type I error rate was computed as the ratio of the number of independently working examinees detected divided by 5,000, the total number of independently working examinees.

The power of the collusion model was studied by running the model on the 100 collusion groups. Within each collusion group, we computed the number of individuals who were detected. Power was computed as the percentage of simulated colluders who were detected by the model.

In addition, we define two terms, *Cluster Type I error rate* and *Cluster Power*, which aim to understand the impact of true and false positives on the clustering structure. Cluster Type I error rate is the percentage of groups in which the number of falsely detected individuals exceeded various thresholds (e.g., the detected group included at least 2 false positives). Cluster Power was computed as the percentage of groups in which the number of detected individuals exceeded various thresholds (e.g., at least five people from the group were detected).

Within each condition, we also examined the cluster integrity. Whereas the power analyses were intended to investigate the extent to which known collusion groups were recovered, the purpose of the cluster integrity analysis was to investigate the extent to which recovered clusters would be interpretable. It stands to reason that the ability to meaningfully interpret recovered clusters is impacted by two variables. First, interpretability is improved when a cluster identifies examinees from a small number of groups and is best when all recovered examinees are from a single group. Second, interpretability suffers if a cluster is too small. A minimum number of examinees must exist to discern a pattern. In the event that a cluster involves examinees from multiple groups, the ability

to interpret clusters will suffer unless each group has enough examinees detected to realize that multiple groups are involved. We examined cluster integrity graphically by studying the number of examinees from each collusion group clustered together.

RESULTS

Type I Error Rate

Although the independent variables in this study all related to different amounts and magnitudes of collusion, because the 5,000 uncontaminated examinees were analyzed along with all contaminated examinees, it was conceivable that we would find that the Type I error rate was affected by different levels of the independent variables. Nevertheless, the results were quite similar across conditions.

Table 6.2 reports the observed Type I error rates from each of the 27 conditions, along with the number of clusters that included 1, 2, 3, or 4 or more uncontaminated

Table 6.2 Empirical Type I Error Rates

<i>G</i>	<i>m</i>	ϕ	Size of Type I Clusters				Total Clusters Recovered*	Type I Error Rate
			1	2	3	4+		
5	25	2.0	15	61	7	2	89	0.0342
		2.5	18	57	5	1		0.0302
		3.0	13	59	8	0		0.0310
	35	2.0	15	49	5	0	82	0.0266
		2.5	10	59	13	1		0.0342
		3.0	13	54	6	2		0.0294
	50	2.0	16	56	6	1	138	0.0312
		2.5	21	47	7	2		0.0298
		3.0	11	61	6	3		0.0364
10	25	2.0	23	47	3	1	93	0.0280
		2.5	31	49	5	0		0.0300
		3.0	20	50	5	3		0.0294
	35	2.0	27	46	3	1	125	0.0264
		2.5	27	43	12	0		0.0298
		3.0	22	67	6	1		0.0368
	50	2.0	19	52	5	1	160	0.0306
		2.5	18	53	8	3		0.0320
		3.0	7	52	5	3		0.0296
20	25	2.0	44	50	3	1	156	0.0314
		2.5	44	56	2	0		0.0324
		3.0	23	56	6	1		0.0326
	35	2.0	36	50	2	0	218	0.0284
		2.5	24	35	4	2		0.0238
		3.0	22	45	3	2		0.0274
	50	2.0	26	45	6	1	157	0.0322
		2.5	7	41	3	0		0.0230
		3.0	16	48	9	2		0.0314

*Note: The total number of clusters recovered equals the sum of all the clusters including Type I errors and all clusters including only colluders.

Table 6.3 Empirical Cluster Type I Error Rates

Independent Variables	Size of Type I Clusters				
	0	1	2	3	4+
$G = 5$	0.368	0.117	0.448	0.056	0.011
$G = 10$	0.462	0.145	0.344	0.039	0.010
$G = 20$	0.532	0.158	0.279	0.025	0.006
$m = 25$	0.403	0.179	0.376	0.034	0.007
$m = 35$	0.484	0.143	0.327	0.039	0.007
$m = 50$	0.498	0.106	0.343	0.041	0.012
$\phi = 2.0$	0.405	0.181	0.374	0.033	0.007
$\phi = 2.5$	0.473	0.149	0.328	0.044	0.007
$\phi = 3.0$	0.502	0.103	0.345	0.038	0.012
Average	0.463	0.142	0.348	0.038	0.009

examinees, and the total number of clusters recovered, regardless of whether those clusters included contaminated or uncontaminated examinees. In all circumstances studied, the Type I error rate was well controlled, if not slightly conservative. Over all conditions, the Type I error rates fell between 0.0238 and 0.0368. The average Type I error rate was 0.0303, and the standard deviation across conditions was just .0032. Type I error rates did not appear to be affected by group size, the number of contaminated items, or the contamination effect.

The Type I errors were distributed across clusters such that the vast majority of recovered clusters that included at least one uncontaminated individual actually included no more than two such examinees. Table 6.3 summarizes the data from Table 6.2 to better illustrate the Cluster Type I error rates. Overall, 46.3% of the clusters included no examinees who were falsely detected. The percentage of clusters free of uncontaminated examinees increased as a function of group size, the percentage of contaminated items, and contamination effect. As one can see from Table 6.3, the vast majority of recovered clusters included zero, one, or two uncontaminated examinees. Only 3.8% of the recovered clusters included three Type I errors, and 0.9% of the clusters included as many as four uncontaminated individuals (i.e., across all conditions, the average number of clusters with at least three falsely detected individuals was 4.7%). Across the 27 conditions, the percentage of total clusters that included at least three uncontaminated examinees never exceeded 10.1%, and the percentage including four or more never exceeded 2.2% (both for the first condition presented in Table 6.2). Therefore, it appears as though the Cluster Type I error rate was generally well controlled, though the small group size \times less compromise \times smaller contamination effects conditions were more susceptible to Type I error contamination. Furthermore, the nature of the way in which the Type I errors manifested themselves suggests that if practitioners limit themselves to interpreting only those clusters that are sufficiently large (e.g., those containing five or more examinees), the likelihood of Type I errors having a serious impact is fairly minimal. This can be seen more clearly by examining Figures 6.7–6.9, which will be discussed more fully within the Cluster Integrity section.

Power

Figure 6.3 shows the power curves for each of the nine $G = 5$ (solid symbol) and $G = 20$ (open symbol) conditions studied. Power for the $G = 10$ conditions are not presented on this graph because (a) group size had the smallest effect of all independent variables, (b) those curves tended to fall in between the curves for $G = 5$ and $G = 20$, (c) power data when $G = 10$ are provided in Figures 6.4–6.6, and (d) plotting another nine graphs on this set of axes would make it too difficult to discern patterns.

From Figure 6.3, three patterns readily emerge. First, the variable having the largest impact on detection is the number of contaminated items (shown as a dotted line for $m = 25$, as a solid line for $m = 35$, and as a dashed line for $m = 50$). On a 50-item test, when only 25 items were contaminated, the detection rates were relatively low, even for large values of ϕ (shown as a circle for $\phi = 2.0$, as a triangle for $\phi = 2.5$, and as a square for $\phi = 3.0$). Similarly, when all items were contaminated, the detection rates were quite high, even for small values of ϕ . The second clear pattern is that the magnitude of ϕ correlated positively with detection rates. Holding constant the group size and number of contaminated items, power increased uniformly as ϕ increased from 2.0 to 2.5 to 3.0. Finally, one can see that group size had only a small effect on the relative

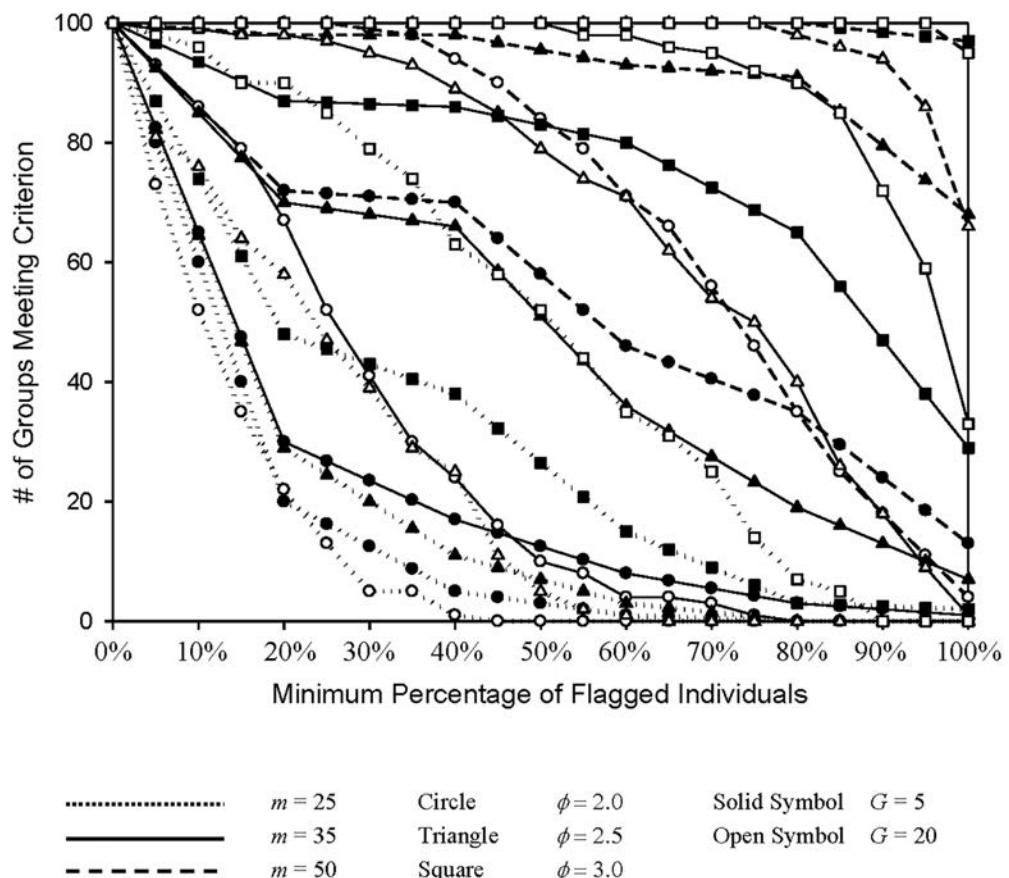


Figure 6.3 Number of Groups in Which a Minimum Percentage of Individuals Were Flagged for Simulated Collusion Groups of Size 5 or 20

ordering of the conditions, and the differences in power as a function of group size were often rather small, especially for the most and least powerful conditions. Where differences did exist, the power for large groups was typically greater, although there are some instances where the curves for $G = 5$ and $G = 20$ cross so that the smaller group yields higher power to detect certain percentages of its members. Because the pattern of results was similar across the three group sizes, in the interest of space, only results for the $G = 10$ conditions will be presented in greater depth.

Figures 6.4 through 6.6 provide visual representations of the power of this method to detect groups of interrelated examinees when $G = 10$, and the number of contaminated items is 25, 35, and 50, respectively. Within each figure, three graphs are presented, one each for the three levels of ϕ . The graphs themselves show a stacked vertical bar for each of the 100 known collusion groups. Each individual stack presents the number of

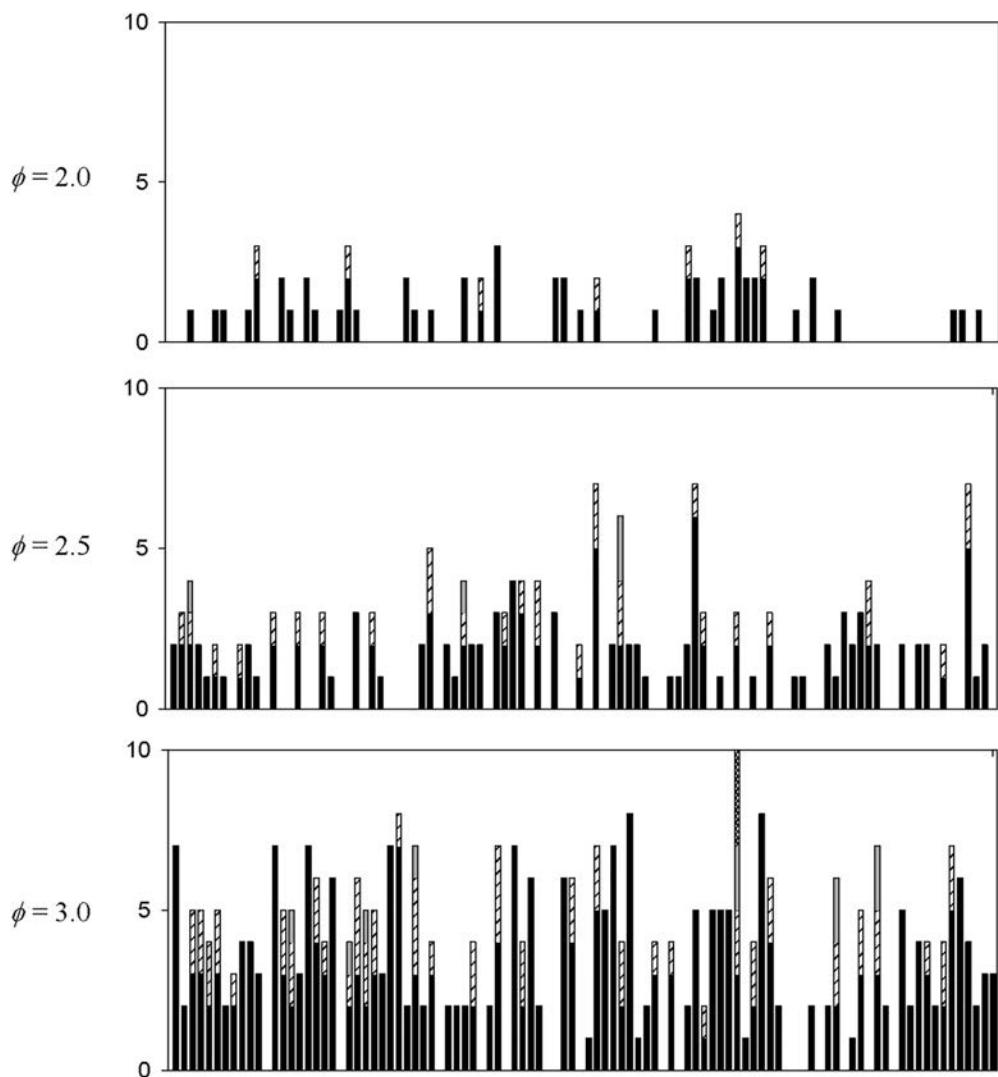


Figure 6.4 Pattern of Examinees Clustering Within Each Contaminated Group When $G = 10$ and $m = 25$

examinees from that particular collusion group that were identified as belonging to a common cluster, and the number of stacks within each bar shows the number of clusters into which members from that collusion group were sorted. The overall height of each bar (across all stacks) corresponds to the total number of examinees in each group who were detected as anomalous (out of a possible 10) and provide the data parallel to what is summarized for the $G = 5$ and $G = 20$ conditions within Figure 6.3.

From Figures 6.4–6.6, it is clear that the quality of detection improves markedly as m and ϕ increase. As m and ϕ increase, not only does the overall number of contaminated examinees detected increase (i.e., the graphs contain less white space) but so too does the tendency for those examinees to be detected as belonging to a common cluster (i.e., bars are mostly a single shade). As an example of this, in the three most powerful

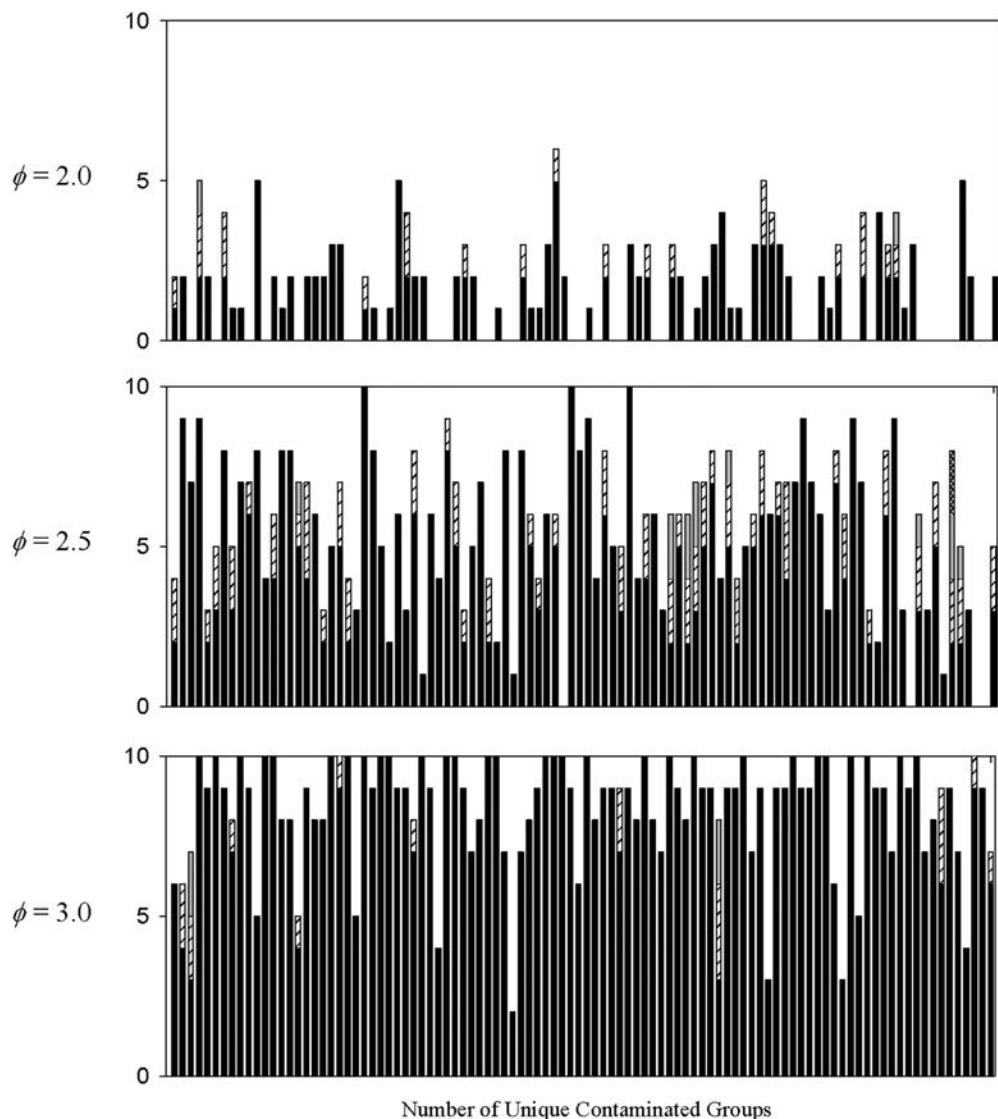


Figure 6.5 Pattern of Examinees Clustering Within Each Contaminated Group When $G = 10$ and $m = 35$

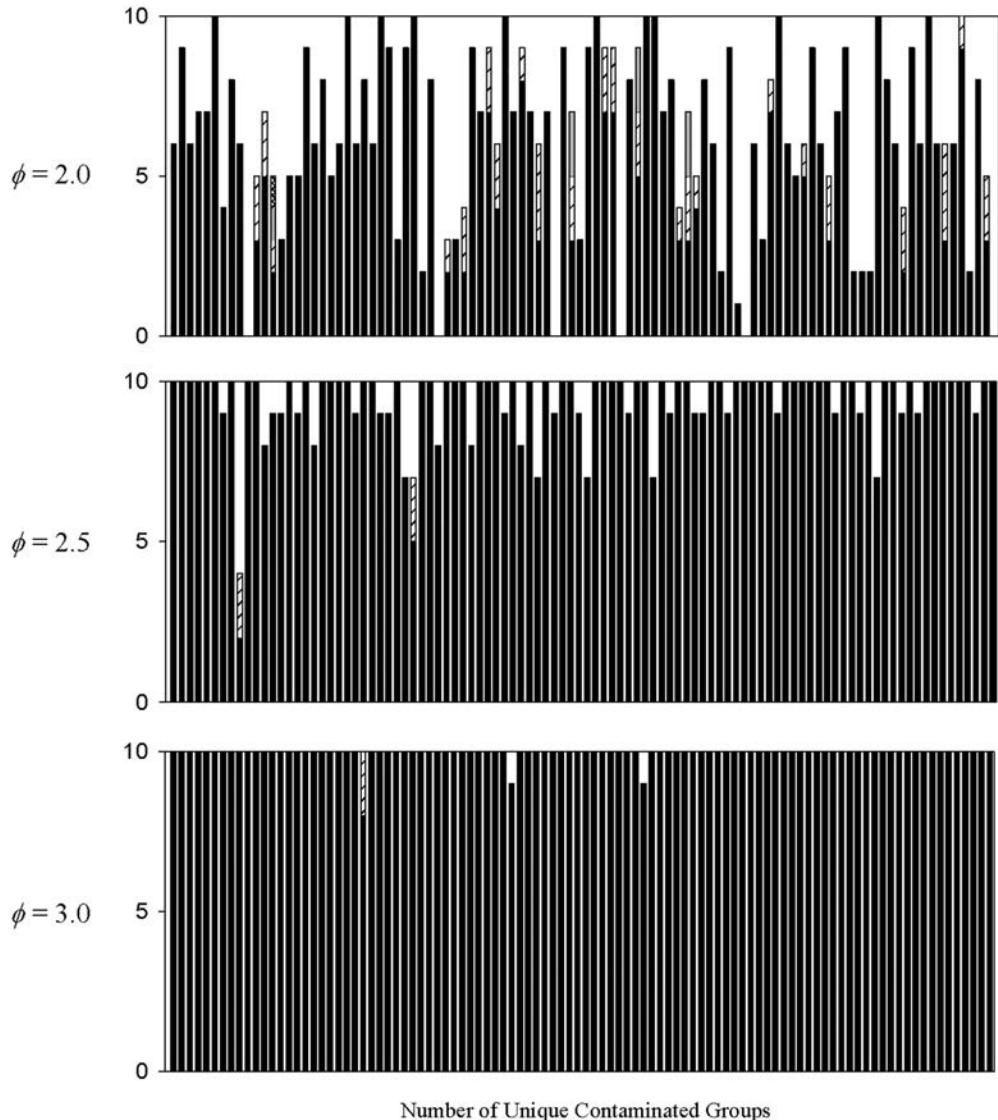


Figure 6.6 Pattern of Examinees Clustering Within Each Contaminated Group When $G = 10$ and $m = 5$

conditions ($\phi = 3.0 \times m = 35$, $\phi = 2.5 \times m = 50$, and $\phi = 3.0 \times m = 50$), only 11%, 2%, and 1% of the collusion groups, respectively, required more than one cluster to recover its members, whereas the percentages were substantially higher for the other conditions (with the exception of the $\phi = 2.0 \times m = 25$ condition, which had a low percentage simply because the probability of anyone being detected in this condition was so low).

It is anticipated that, in practice, interpretation will be facilitated by having many group members identified in a common cluster (or perhaps two medium-sized clusters), as opposed to scattering the identified individuals across three or more clusters. From Figures 6.4–6.6, it appears as though the higher the collusion severity parameter, the more interpretable the clusters will be. This issue is looked at more closely in the cluster integrity analyses that follow.

Cluster Integrity

In evaluating the effectiveness of this procedure, it is important to consider not only the true and false detection rates but also the extent to which the solutions are interpretable and are likely to result in successful investigations. As such, it is necessary to look at the individually identified clusters, and whether they are likely to provide enough information to be helpful during an investigation. Figures 6.7–6.9 provide visual representations of what we will call cluster integrity, a visual description of the composition of detected clusters, when $G = 10$, but m and ϕ vary.

Along the horizontal axis of these figures are stacked bar graphs for each of the detected clusters. The number of groups stacked together within a cluster corresponds

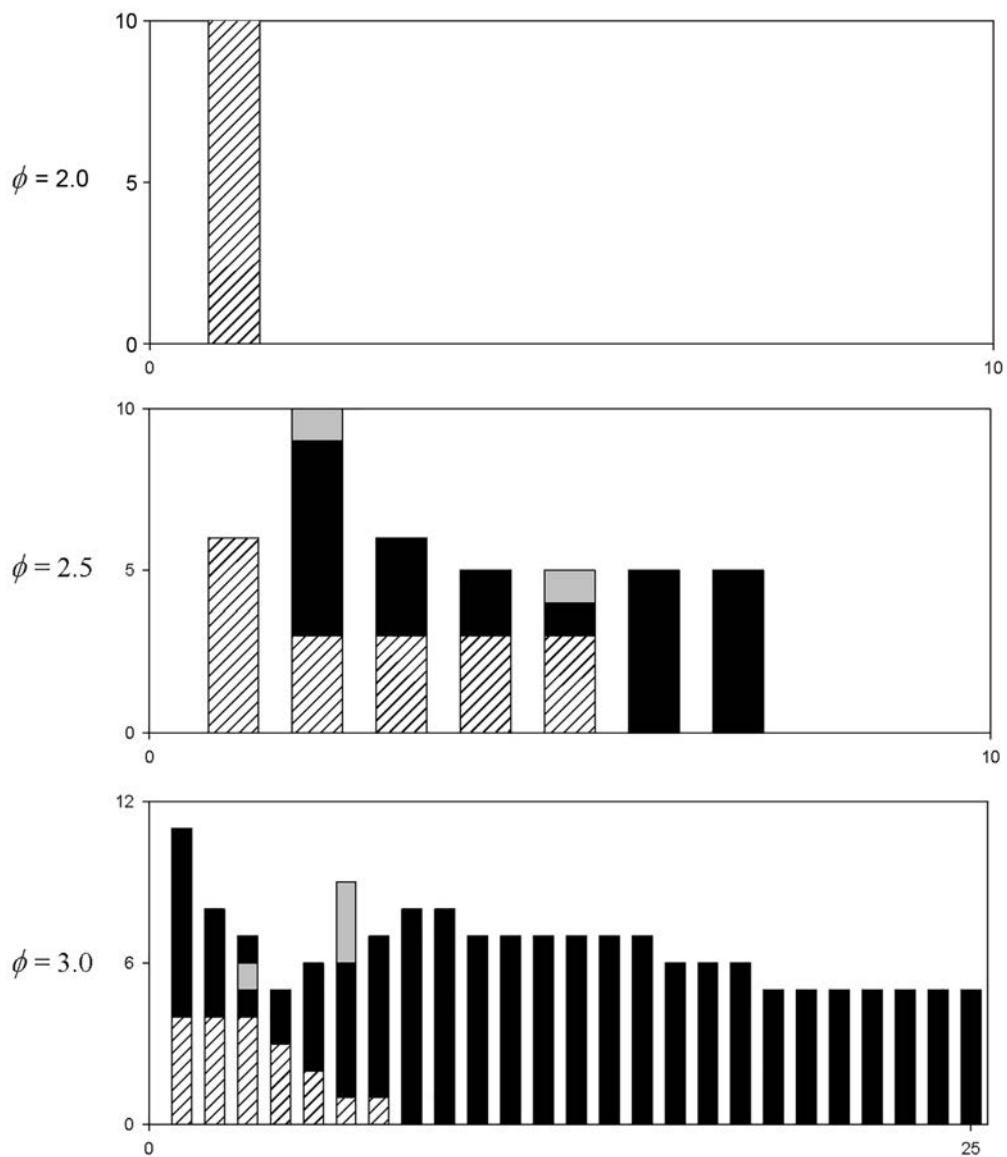


Figure 6.7 Cluster Integrity Pattern (for clusters with >5 examinees) When $G = 10$ and $m = 25$

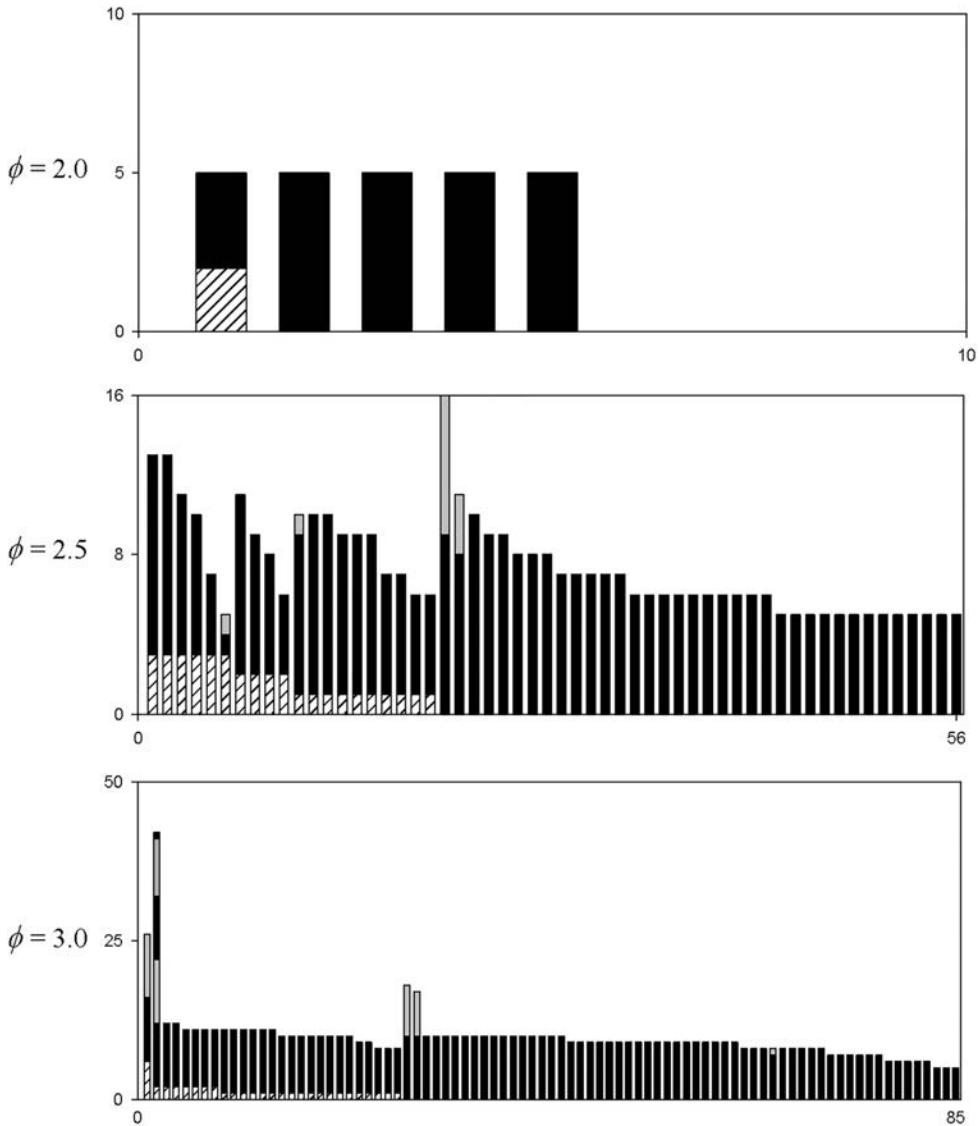


Figure 6.8 Cluster Integrity Pattern (for clusters with >5 examinees) When $G = 10$ and $m = 35$

to the number of unique collusion groups that were clustered together by this procedure, and the height of each stack corresponds to the number of individuals within each of those collusion groups. For ease of interpretation, the recovered clusters in Figures 6.7–6.9 have been organized as follows. Clusters that include falsely detected examinees (Type I errors) are presented first and in descending order of the number of false positives. All Type I errors are presented in white bars with diagonal line shading. Clusters with only colluders are presented next in alternating black and gray bars (again, in descending order of cluster size), with each color change signaling representatives from an additional collusion group. For clusters that contain both Type I errors and true colluders, they are first ordered by the number of Type I errors; clusters with the same number of false positives are ordered by overall cluster size. As an example,

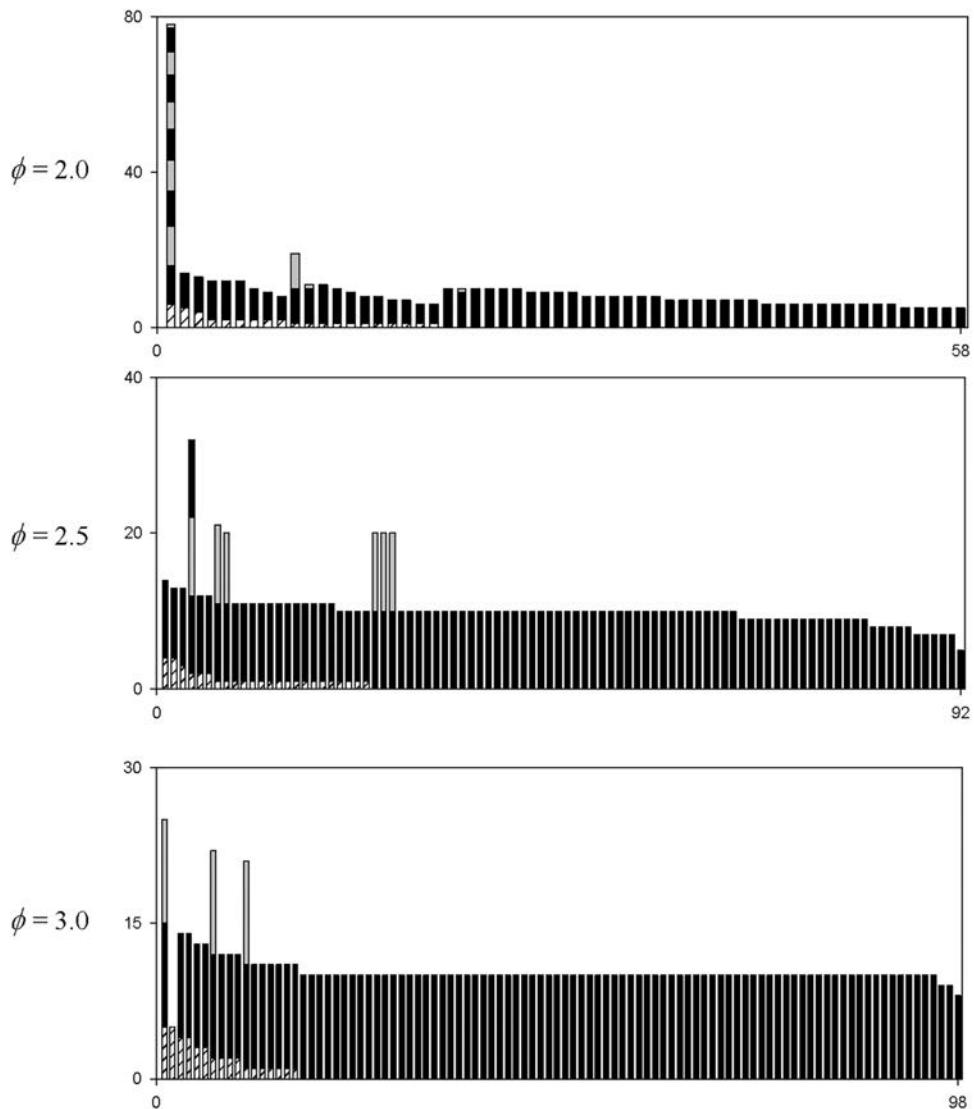


Figure 6.9 Cluster Integrity Pattern (for clusters with \geq examinees) When $G = 10$ and $m = 50$

in the middle graph in Figure 6.7, the second bar shows that this particular recovered cluster included three noncolluding examinees (Type I errors), six examinees from a single collusion group, and one examinee from a different collusion group.

When using this method in practice, the investigator will not be privy to the true group composition. All that an investigator will have at his or her disposal is the detected cluster groupings. Therefore, there are three factors that will greatly contribute to one's being able to correctly interpret a collusion cluster.

- First, the fewer groups clustered together, the better. A cluster that consists only of members from one collusion group will be easier to interpret correctly than a cluster that includes members from multiple groups.

- Second, the more people who are identified from a particular collusion group, the easier it will be to recognize that these individuals share certain background characteristics. This is probably especially true if multiple groups are clustered together. From an interpretability perspective, there is a big difference between clustering together 2–3 members of each of 10 groups, versus linking together 10 nearly intact groups of 10–20 examinees each.
- Finally, the cluster must be large enough that we are confident that the cluster is real. As shown with the Type I error data, given the criteria used to define clusters, it is not particularly uncommon for two or three examinees to be clustered together, even if none of the examinees involved was engaged in any test collusion. As a result of this criterion, and to facilitate seeing patterns in the data, Figures 6.7–6.9 display only recovered clusters with at least five examinees. Suffice it to say that for all conditions, there were also many recovered clusters of size 2, 3, or 4 that are not shown.

With regard to the first criterion above, overwhelmingly, the clusters recovered examinees from one or two collusion groups. Across the nine conditions shown in Figures 6.7–6.9, only three *super-clusters* were created in which clusters combined individuals from three or more collusion groups. The bottom-most graph in Figure 6.7 shows one cluster that combined one member from each of three different collusion groups (along with some Type I errors). As there is very little meaningful data tying together these examinees, it is unlikely that this cluster would be able to be interpreted. In the other two instances of multiple collusion groups clustering together, a total of five (bottom graph of Figure 6.8) or 10 (top graph of Figure 6.9) groups were combined. Although these large and heterogeneous clusters may make interpretation more challenging, it is encouraging that these super-clusters were very good with respect to the second criterion. That is, they appeared to consist of many largely intact groups (i.e., groups where the majority of the group members were identified) rather than many very small groups. As an example, within the super-cluster in the $m = 50 \times \phi = 2.0$ condition, at least six (of the 10) group members were identified for 9 of the 10 groups. Similarly, within the super-cluster in the $m = 35 \times \phi = 3.0$ condition, for four of the five groups clustered together, at least nine group members were identified.

The various graphs also differ considerably with respect to the third criterion that they must produce large clusters. Generally speaking, the endpoint of the x -axes in Figures 6.7–6.9 show the number of clusters in each condition with at least five examinees. For the topmost two graphs in Figure 6.7 and the topmost graph in Figure 6.8, however, the x -axis was fixed at 10 to improve the presentation of the data. As can be seen, in each of the four least less powerful conditions (i.e., all three $m = 25$ conditions and the $m = 35 \times \phi = 2.0$ condition), at most 25 clusters contained five or more people, and in three of those conditions, no more than seven clusters were sufficiently large to present. Also, in the three least powerful conditions, the majority of those clusters were sufficiently contaminated with Type I errors that interpretation would have been challenging. In contrast, in the two conditions with moderate power (i.e., $m = 35 \times \phi = 2.5$, and $m = 50 \times \phi = 2.0$) 56–58 clusters included at least five examinees. In the three high-power conditions ($m = 35 \times \phi = 3.0$, $m = 50 \times \phi = 2.5$, and $m = 50 \times \phi = 3.0$), between 85–98 clusters identified at least five examinees, and nearly all of them identified at least eight examinees.

Although Figures 6.4–6.6 reported cluster integrity when group size was fixed at 10, the overall pattern across different group sizes was also in line with expectations. In

particular, as group size increased, the number (and percentage) of clusters with five or more people from a common cluster increased, as did the tendency to produce super-clusters by linking together many intact groups.

Finally, it is worth reiterating here that the cluster integrity graphs show that the Type I errors do not appear to interfere with the clustering of known collusion groups. Furthermore, among the larger clusters—the ones most likely to be investigated—the Type I errors also do not contaminate the clusters to a sufficient degree that it is likely to impede an investigation.

SIMULATION STUDY DISCUSSION

This study shows that, under realistic circumstances, it is possible to recover clusters of interrelated examinees that would likely be very helpful during an investigation, provided the amount and magnitude of collusion is reasonably high—precisely the circumstances in which it is most important to detect collusion. Even under realistic conditions in which contaminated and uncontaminated individuals are combined in the same dataset, the overall Type I error rate of this procedure's methodology remains well controlled. Furthermore, because such a high percentage of the Type I errors appeared in clusters of one, two, or three, in a practical setting where it is likely that only moderate-to-large clusters will be investigated further, the overall impact of Type I errors on cluster interpretation figures to be minimal. In addition, for moderate-to-large amounts of collusion, the clustering technique presented here was generally successful at recovering known collusion groups, both with regard to the overall numbers of colluding individuals detected and in recovering a collusion pattern that will be interpretable during an investigation. Although there were a few instances in which super-clusters were formed, the integrity of those clusters was still quite high, as each super-cluster included many individuals from each collusion group identified.

APPLICATION TO COMMON CREDENTIALING DATASET

Because collusion detection would be applicable to any type of security breach which would produce groups of examinees with similar answer patterns, it would have been appropriate for us to apply the method to either the common K-12 dataset—to look for signs of illegal coaching or test tampering—or to the common credentialing dataset to look for evidence of preknowledge. We ultimately decided to apply the method to the credentialing dataset because that dataset includes known compromise, whereas the K-12 dataset does not.

Method

The clustering method was run separately by form. Because it was not expected that examinees would have preknowledge of pilot items, analyses were conducted using only the 170 operational items. In running the analysis, we first computed $M4$ among all possible pairs sharing a test form. This resulted in a total of 1,337,430 pairs completing Form 1 and 1,350,546 pairs completing Form 2. To control for the large number of indexes computed for each examinee (any one of which could result in the examinee being flagged as anomalous), a multiple comparisons adjustment was used to control the false positive rate at the examinee level. As discussed earlier, because the individual, uncorrected $M4$ is computed as $M4 = -\log_{10}(p)$, where p is the probability associated

with the generalized trinomial, the multiple comparisons adjustment was incorporated by computing $M4 = -\log_{10}(1 - (1 - p)^{(N-1)/2})$, where N is the number of examinees taking the test form. For consistency with the simulation study, $M4$ values greater than 1.301 (i.e., an $\alpha = .05$ criterion) were used here, though results using $M4$ values greater than 3.0 (i.e., an $\alpha = .001$ criterion) are also provided to see how a more typical false positive rate might affect cluster extraction and interpretability.

Results

With an $\alpha = .05$, we would expect approximately 82 examinees to be falsely detected on each of the two forms. This is obviously a very large number and suggests that, in practice, we would likely want to use a smaller α level. However, given that our simulation showed that even with an $\alpha = .05$, Type I errors tended not to dramatically affect the interpretation of larger clusters, while allowing more true cheaters to be detected, we wanted to start by looking at those results. A total of 284 individuals were identified, including 134 on Form 1 and 150 on Form 2. Given $M4$'s strong (if not conservative) control of the false positive rate (Maynes, this volume), this clearly suggests that a lot of examinees were engaged in some type of collusion during the exam. Examinees were grouped into 90 clusters. However, 84 of those clusters were of size 4 or smaller. The remaining six clusters—three on each form—ranged in size from six to 27 and included a total of 94 candidates. Given the large number of expected Type I errors, it was decided to only interpret clusters with at least five candidates. Because the number of identified candidates is so high, we will attempt to describe the characteristics of these groups, rather than list and describe each flagged candidate.

The largest cluster for Form 1 included 27 candidates. Of these, 21 were from India. Of the remaining six candidates not from India, they came from five different countries, sought licensure in five different states, and tested in four different states, with one state being common to three of them. However, that one common state was also the largest state. Of the 21 from India, however, nine sought licensure in one state, while the remaining 12 sought licensure in a second state. There were 11 candidates who tested in one of two states; the population of candidates tested in these two states just 12.2% of the time. Six of the candidates tested at the same site. Of the 21 candidates from India, 11 were flagged by the test sponsor; none of the six candidates from other countries were flagged.

The second largest Form 1 cluster included 18 candidates. Twelve candidates were from the Philippines; three were from India. Eight candidates sought licensure from the same state, again, the largest state in the dataset. Three candidates from the Philippines attended the same school; two attended another. The remaining candidates all attended different schools. No two candidates tested at the same site. Only one of these examinees—one from India—was flagged by the test sponsor.

The third Form 1 cluster included only six candidates. Five were from the Philippines; one was from India. They all attended different schools, applied for licensure in six different states, and sat for the exam in different sites (and in different states). None of these six were flagged by the test sponsor.

The largest cluster for Form 2 involved 19 candidates. All candidates were from India. Ten candidates applied for licensure in one state; eight applied in a second. There was a single institution at which seven of these candidates attended school. There were three others which were attended by two candidates. Five candidates took the exam at the same facility. A total of 11 candidates in this cluster were flagged by the test sponsor.

The last two clusters were both of size 12 and consisted almost entirely of candidates from the United States. Within clusters, there was nothing noteworthy to report with respect to unusual similarity among any of the background variables. None of the 24 individuals in these two clusters was suspected of misconduct by the test sponsor.

All six of these clusters were suspicious to some degree. It seems unlikely to be a purely chance finding that 56% of the detected examinees were either from India or the Philippines, when these two countries only represented approximately 19% of all candidates. All six of the clusters were surprisingly homogeneous with respect to country. Within the clusters, it was often the case that certain licensure states, test sites, or schools were overrepresented; however, except for a few isolated instances, the extent of overrepresentation wasn't dramatic. And with all clusters, there appeared to be examples of cases that did not fit with the others, perhaps suggesting that the clusters were a bit too contaminated with Type I errors, or perhaps suggesting that the $\alpha = .05$ criterion simply made it too easy for tangentially related clusters to combine into a super-cluster.

The same analysis was repeated using an $\alpha = .001$ criterion to identify significant linkages between individuals/clusters. Given the sample sizes and the manner in which we controlled for multiple comparisons, we would expect between one and two individuals per form to be flagged erroneously (i.e., 0 – 1 significant *M*₄ indexes among noncolluding examinees). Under this criterion, a total of 14 Form 1 examinees and 11 Form 2 examinees were identified, clearly more than would be expected due to chance alone. Identified examinees were grouped into five clusters, two for Form 1 and three for Form 2. Of the five clusters, three were clusters of two (i.e., a pair of examinees). Although it is entirely possible that two examinees were working together or engaged in a traditional answer copying paradigm (although we have no seating chart information that could help us with such a hypothesis), in light of the simulation results showing the propensity for clusters of two to be Type I errors, those clusters will not be interpreted here.

Table 6.4 shows the two clusters with five or more examinees, one for each form, along with information on the country of origin, state in which the candidate applied for licensure, school in which the candidate was trained, the site at which the test was administered, and the state in which the test was administered. A field is also presented indicating whether these particular candidates were believed to have engaged in inappropriate testing activities by the test sponsor. Several interesting patterns are readily apparent. First, 18 of the 19 individuals involved in these clusters are from India. Considering that only 320 of the 3,280 (9.8%) candidates sitting for this exam were from India, this is an extraordinarily unlikely event (less than 1 in 90 quadrillion, based on a binomial). In addition, all of the candidates, including the one from Egypt, were applying for licensure in one of two states. There is also considerable similarity with respect to the locations of the testing sites. Unsurprisingly, the state in which the candidate tested is often the same as the state in which the candidate is applying for licensure. In the entire dataset, this phenomenon occurred in 74.5% of the cases. However, here, 8 of the 19 applied for licensure in a different state than they tested in; a result as extreme as this is expected only 0.27% of the time. Furthermore, when one examines the specific states in which candidates tested, they again find a good deal of similarity. Finally, although there was plenty of variability in school code and test site, some patterns were still evident. For example, across the two forms, there were five candidates who attended School 8198 and six who attended School 5530. Also, nine of the candidates tested at Site 2305.

Table 6.4 Clusters Extracted with $\alpha = .001$

Cluster ID	Flagged	Country	State	School	Test Site	Site State
Form 1 Clusters						
1	0	Egypt	42	5447	5203	42
	1	India	42	5530	2305	28
	1	India	42	8092	5862	54
	0	India	42	8113	5203	42
	1	India	42	8119	5880	42
	1	India	42	8155	5303	54
	1	India	42	8198	5302	54
	1	India	42	8198	1437	43
	1	India	28	5530	2305	28
	1	India	28	8152	2305	28
	1	India	28	8172	2305	28
	1	India	28	8198	2305	28
Form 2 Clusters						
2	1	India	28	5530	1	8
	1	India	28	5530	2305	28
	1	India	28	5530	2331	28
	1	India	28	8172	2305	28
	1	India	28	8198	2305	28
	1	India	42	5530	47	8
	0	India	42	8198	2305	28

Adding to the evidence that these clusters really are candidates who were engaged in collusion is the fact that 16 of the 19 candidates were also flagged by the test sponsor at the conclusion of their security investigation. Of the three who were not, one is the candidate from Egypt. This candidate does overlap with many of the others with respect to where they applied for licensure and the state in which they tested and even tested at the same test site as another candidate on this list (who, interestingly, is the only other Form 1 individual not flagged by the test sponsor). The one candidate from Form 2 who was not identified by the test sponsor attended school with and tested at the same site as many of the others and so should certainly be viewed with extra suspicion.

Based on these data alone, not enough is known about the three unflagged candidates appearing in these clusters; however, if the test sponsor had this clustering information at the time of the investigation, they would have been able to include them in the investigation to determine whether there was enough ancillary evidence to suspect misconduct.

Common Data Discussion

Results from the real-data application help illustrate the potential utility of this method in practice. Analyzing results separately by test form and focusing our interpretations solely on clusters with at least five examinees, we found evidence that several groups of examinees may have engaged in collusion. Although the clustering technique was based

solely on the statistical similarity among their test responses, further investigation also found a high degree of similarity with respect to the background characteristics of candidates within a cluster, lending support to the hypothesis that the individuals within a cluster were somehow working together.

As would be expected, the number of clusters and the number of candidates detected were both higher when the $\alpha = .05$ criterion was used. However, the more liberal flagging criterion also resulted in more noise in the data (i.e., Type I errors); as such, the cluster integrity or the ease with which clusters could be interpreted improved under the more conservative criterion. With this in mind, one strategy may be to use a reasonably conservative criterion for purposes of narrowing in on the cluster characteristics, then use a more liberal criterion to identify additional candidates with similar profiles who may warrant some follow-up from the individual conducting the investigation.

Because these are real data, it is not possible to know whether the identified clusters included examinees from one group or multiple groups, nor is it possible to know which of those examinees may have been falsely detected. This dataset is well suited for studying cheating behaviors because it includes the conclusions from the test sponsor about whether or not each specific examinee cheated. Those determinations were reached following a thorough investigation and so draw upon a lot of data that are not available here. However, it is also likely that there were some candidates who were suspicious but for whom the test sponsor did not have quite enough evidence of involvement to warrant invalidating their scores. This is especially true given the paucity of statistical methods available to help programs identify groups of interrelated examinees. Therefore, the fact that some of the candidates detected here were not identified by the test sponsor should not suggest that they should be automatically dismissed as Type I errors. Similarly, the failure of this method to detect a higher percentage of the individuals flagged by the program may also reflect the result of an investigation, in which many nonstatistical pieces of evidence were likely utilized.

CONCLUSION

This study introduced a new approach for identifying groups of examinees who engaged in test collusion. The approach was found to be quite promising in that it works well at identifying high percentages of simulated collusion groups while minimizing the impact of false positives, especially as the extent of collusion increases.

One important finding is that this procedure worked well in spite of estimating parameters on a dataset simulated to include both contaminated and uncontaminated examinees. The inclusion of contaminated examinees will result in item parameter estimates that are less accurate. In particular, it is expected that the parameters will be skewed so that the answer being provided by the “source” is more probable. In many cases, this will be the correct answer, in which case those items will appear easier than they should be. It is reasonable to assume that these spurious item parameters will lead to probabilities of answer matches that are also spuriously high, thereby requiring examinees to share more answer matches in common before being characterized as atypical. As a result, when item parameters need to be estimated from data that are believed to be contaminated, it is likely that the power of the procedure will drop somewhat. If data from more secure administrations can be used for parameter estimation, it will likely facilitate the detection of collusion, should it occur.

If a program cannot rely on data from a previous administration and must estimate item parameters from a dataset believed to be contaminated, one strategy might be to

adopt an iterative approach to parameter estimation. The model first would be fit using all examinees, after which those examinees who were identified within a cluster would be removed from the dataset and the item parameters would be reestimated using the purified dataset of examinees. These purified item parameter estimates could then be used to recompute the similarity indexes between all possible pairs, and the clustering algorithm reapplied.

This study did not manipulate the number of colluding examinees. However, it stands to reason that as the number of contaminated examinees increases, holding all else constant, the power to detect them will decrease. The multiple comparisons approach to control the probability of falsely identifying a candidate is affected by the total sample size, even though some of those candidates cannot be falsely detected (because they are colluders). Therefore, as more contaminated examinees are embedded in the dataset, the critical value for detection becomes spuriously high, which will lower the overall detection rate.

One approach that might be useful to improve the power is to estimate item parameters using all examinees available but analyze clusters in smaller batches, based on where evidence for collusion will be most apparent. This strategy was adopted here with the application to the common licensure dataset, in that Forms 1 and 2 data were analyzed separately, thereby reducing by half the number of comparisons being analyzed and lowering the effective critical value. Another possibility would have been to analyze separately by home country or school. This approach will result in considerably more power to detect clusters within the groups examined and will dramatically reduce the number of Type I errors but will also result in an inability to detect collusion between the groups. Given that many types of collusion do not require a geographic connection, rather a social one, the decision to analyze clusters in smaller batches should be made cautiously.

The clustering method used in this study, the nearest neighbor/single linkage method, has the advantage of being very simple computationally, but does occasionally link together many clusters, even if the majority of elements from each of the clusters are dissimilar. This finding is not unique to this study; indeed, single linkage is well known to have a tendency to chain together cases (Aldenderfer & Blashfield, 1984). Other clustering approaches such as average linkage or complete linkage are more apt to produce distinct clusters. Therefore, utilizing this collusion detection framework with one of these alternative clustering approaches, or using these alternative clustering approaches as a secondary analysis to aid with the interpretation of extra-large clusters, is another area left for future study.

The 0.05α level that was chosen for the simulation study was much more liberal than is customarily used in operational test security work. This was an intentional choice, so as to maximize the opportunities for detection in situations typically encountered in large-scale educational, certification, and licensure settings where exposure rates are often quite high but the amount of dependence is quite low. The overall Type I error rates observed in this study were in keeping with the high α level used here; however, it is important to note that the clustering patterns of Type I errors were such that the overall impact of the Type I errors was only minimal.

Although a number of sophisticated tools exist for identifying pairs of examinees whose answer patterns are unusually alike, comparable resources are lacking to identify multiple examinees producing similar responses. It is our sincere hope that the approach presented here, along with the recommendations from this study, will pave the way towards further research in this area, as testing organizations continue to battle a new generation of test villains to preserve the integrity of their tests.

NOTES

1. This last activity is actually tampering, but when the tampering is performed on several answer sheets consistently, it can be detected using collusion analysis.
2. If Ω is to be controlled at the experiment-wide level (i.e., for all pairs) the adjustment should be based on all pairs $(N^2 - N)/2$ (Wesolowsky, 2000).
3. Wollack and Maynes (2011) demonstrated the proof of concept for this approach by studying the ideal situation where only data from uncontaminated individuals were used to estimate model parameters.

REFERENCES

- Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Newbury Park, CA: Sage.
- Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing*, 2(3), 37–58.
- Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, 40(2), 93–97.
- Belov, D. I., & Armstrong, R. D. (2009). *Detection of answer copying via Kullback-Leibler divergence and K-index (Research Report 09-01)*. Newtown, PA: Law School Admission Council, Inc.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 46, 443–459.
- Chang, H. H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211–222.
- Chen, S.-Y., & Lei, P.-W. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement*, 29, 204–217.
- Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Doong, S. H. (2009). A knowledge-based approach for item exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 34, 530–558.
- Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, 6, 152–165.
- Impara, J. C., Kingsbury, G., Maynes, D., & Fitzgerald, C. (2005, April). *Detecting cheating in computer adaptive tests using data forensics*. Paper presented at the annual meeting of the National Council on Measurement in Education.
- Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, 118, 843–877.
- Leung, C., Chang, H., & Hau, K. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson-Hetter algorithm. *Applied Psychological Measurement*, 26, 376–392.
- Maynes, D. D. (2009, April). *Combining statistical evidence for increased power in detecting cheating*. Presented at the annual conference for the National Council on Measurement in Education, San Diego, CA.
- Maynes, D. (2013). Security among Teachers and Administrators. In J. A. Wollack & J. J. Fremer (Eds.) *Handbook of Test Security* (pp. 173–199). New York, NY: Routledge.
- Maynes, D. D. (2014). Detection of non-independent test taking by similarity analysis. In N. M. Kingston & A. K. Clark (Eds.) *Test Fraud: Statistical Detection and Methodology* (pp. 53–82). New York, NY: Routledge.
- Maynes, D. D. (this volume). Detecting potential collusion among individual examinees using similarity analysis. In G. J. Cizek & J. A. Wollack (Eds.) *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.
- McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement*, 23, 147–160.
- Meijer, R. R., & Sotaridona, L. S. (2006). *Detection of advance item knowledge using response times in computer adaptive testing (Research Report 03-03)*. Newtown, PA: Law School Admission Council, Inc.
- Nering, M. L., Davey, T., & Thompson, T. (1998, July). *A hybrid method for controlling item exposure in computerized adaptive testing*. Paper presented at the annual meeting of the Psychometric Society, Urbana, IL.
- Shu, Z., Henson, R., & Leucht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item comprise. *Psychometrika*, 78, 481–497.
- Sotaridona, L. S., & Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement*, 40, 53–69.
- Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 23, 57–75.

- Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277–292.
- van der Linden, W. J. (2003). Some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 28*, 249–265.
- van der Linden, W. J. (2009). A bivariate lognormal response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics, 34*, 378–394.
- van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*, 365–384.
- van der Linden, W. J., & Sotaridona L. S. (2004). A statistical test for detecting answer copying on multiple-choice tests. *Journal of Educational Measurement, 41*, 361–377.
- van der Linden, W. J., & Sotaridona L. S. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics, 31*, 283–304.
- van der Linden W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika, 68*, 251–265.
- van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*, 199–218.
- Wesolowsky, G. O. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics, 27*, 909–921.
- Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement, 21*, 307–320.
- Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and ability parameters. *Applied Psychological Measurement, 22*(2), 144–152.
- Wollack, J. A., Cohen, A. S., & Serlin, R. C. (2001). Defining error rates and power for detecting answer copying. *Applied Psychological Measurement, 25*(4), 385–404.
- Wollack, J. A., & Maynes, D. (2011, April). *Detection of test collusion using item response data*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Yi, Q., & Chang, H.-H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology, 56*, 359–378.
- Zamost, S., Griffin, D., & Ansari, A. (2012). Exclusive: Doctors cheating on exams. CNN. Retrieved December 28, 2015 from www.cnn.com/2012/01/13/health/prescription-for-cheating/.

7

DETECTING CANDIDATE PREKNOWLEDGE AND COMPROMISED CONTENT USING DIFFERENTIAL PERSON AND ITEM FUNCTIONING

Lisa S. O'Leary and Russell W. Smith

INTRODUCTION

Item compromise is a consistent threat to the validity of certification examinations due to prevalent piracy practices that include the regular theft and unauthorized release of individual items, entire item pools, and exam forms. It has been documented that item compromise is widespread within information technology certification exams in particular; live exam content and items are routinely being exposed on the Internet. The three basic outlets for legitimate and illegitimate advice and content are: exam preparation sites, Internet auction sites, and braindump sites that either formally sell stolen content or informally encourage candidates to share their own recollections about particular certification exams (Smith, 2004; Foster, 2013). According to research (Maynes, 2009) within some high-volume certification testing programs, a majority of candidates (i.e., 85% or more) may have acquired prior item knowledge by purchasing content through braindump sites.

As a result of this problem, certification testing programs have been plagued by skepticism about the legitimacy of candidate exam results and resulting inferences about individuals' knowledge, skills, and abilities. Therefore, this chapter will focus specifically on addressing and diminishing the influence of test fraud, which is defined as "any behavior that inappropriately or illegally captures test questions and/or answers" (Foster, 2013, p. 47), as opposed to other cheating behaviors such as collusion or tampering. Test fraud is prominent in certification programs for a multitude of reasons, including, but not limited to, the high professional stakes linked to successful certification; advances in technology; the computer-based, continual delivery of many of these exams; and the candidates' familiarity and years of experience with technology (Smith, 2004; Wollack & Fremer, 2013). Given the prevalence of test fraud—particularly piracy—in certification testing programs, managing item compromise is a top priority regarding exam security.

Test fraud can damage testing programs on several levels. Intangible costs are a loss of credibility and face validity with key stakeholders (including candidates themselves

and employers). Tangible costs are associated with continual efforts geared at item protection and cheating detection, as well as item and test development. Item compromise as a security concern is rampant within certification examinations; it constitutes serious threats to the validity of score interpretation and use. As Impara and Foster (2006) highlight, cheating introduces construct-irrelevant variance; scores may not accurately represent underlying content knowledge but instead “how a particular set of test questions has been answered or tasks performed through inappropriate means” (pp. 91–92). Once items or entire examinations are exposed, candidates’ score integrity is compromised and the validity of the inferences being made based on the exam scores can be questioned (Wollack & Fremer, 2013). Consequently, it becomes difficult to discern if candidate performance is due to true ability or cheating through prior unauthorized access to exam content.

Additionally, item degradation occurs as a result of compromise. Such compromise jeopardizes the quality, utility, and functioning of individual items as well as entire item banks. Item degradation is caused by the deterioration of desirable item characteristics over time, including, but not limited to, a reduction in content relevance and representativeness, loss of quality of technical characteristics (i.e., item difficulty and reliability), and a decrease in utility of the correlation between the item and construct of interest (Yang, Ferdous, & Chin, 2007). Maynes (2013) describes the residual effect of a large number of candidates having access to stolen content as a “three-fold issue” (p. 180) that shifts classical item statistics, results in unexpectedly high levels of performance, and presents evidence of collusion through response pattern similarities. This item degradation threatens the statistical assumptions and psychometric models underlying these certification exams as well as the individual testing outcomes for candidates. Therefore, the ability to detect and mitigate the negative effects of prior item compromise in a certification context where prevention of test theft is not realistic is critically important in maintaining evidence of the validity of these testing programs.

Item compromise negatively impacts item statistics. Han and Hambleton (2008) noted that proactive, systematic efforts should be made to identify exposed items, retire, and replace those items with less exposed (or ideally new) items before the impact on the item statistics is too drastic. While Han and Hambleton (2008) indicate that concerted security efforts should be put in place to initially protect from the test fraud and piracy, they recognize that continuous data forensics efforts are also necessary to assess the extent of compromised content. Therefore, it is important for testing programs with perpetually at-risk exam content to develop procedures around data forensics to quantify the degree of unauthorized released content, as well as policies around item compromise controls to gauge the extent of compromise in item banks and retire and refresh content as necessary (Impara & Foster, 2006).

For the purposes of this chapter, item compromise is operationalized as “the number or percent of people with pre-knowledge of the item before taking the test” as opposed to the more traditional definition of item exposure as the “number of naturally occurring presentations of the questions” (Foster, 2013, p. 79). A solid research foundation has developed in recent years detailing how to detect and statistically control the standard item exposure that can be expected from the use of computer-adaptive testing algorithms and logic to minimize the likelihood of item compromise (Veerkamp & Glas, 2000; Han, 2003; Lu & Hambleton, 2003; Han & Hambleton, 2004). The authors of this research recommend that extensive item banks should be created to enable the regular replacement of items once overexposure and compromise has been detected (Han & Hambleton, 2008; Foster, 2013). While the authors of this paper advocate for

developing large item banks to address issues of item compromise, the focus here will be more on identifying compromised items and replacing those with new items as necessary to allow for the continual administration of some unexposed content amongst largely exposed item pools instead of controlling the routine usage of particular items. Timeliness is of the essence to minimize the impact of item compromise and maximize measurement integrity; extended use of exposed test items increases the extent of sharing and opportunity for prior knowledge of exam content (Carson, 2013). Test fraud can occur immediately following the initial release of exam forms, sometimes within days or even hours (Smith, 2004; Maynes, 2009). Therefore, it has become of utmost importance that new items are continuously produced to support content refreshing and replacement.

Testing programs have had to develop methods through which they can identify exposed items and aberrant candidate response patterns to reduce the influence of test fraud on the validity of their testing programs. Several common practices are used within the testing community to detect irregular testing behavior, such as investigation into item response latencies (Maynes, 2013), comparisons of total exam performance by total exam time (Smith & Davis-Becker, 2011), and score patterns on Trojan Horse items (Maynes, 2009). These practices help detect suspicious candidate behavior (i.e., candidates who answer items quickly but receive high scores and/or consistently answer correctly according to intentionally miskeyed items). However, these techniques are limited in their enforceability. Candidates can adjust their response patterns to avoid detection. Additionally, these methods offer little assistance to identify specific compromised content despite giving some indication of the overall extent of exposure.

Many testing programs have thus incorporated the application of statistical and probability-based psychometric methods into their security analyses to detect cheating on examinations. Indeed, techniques have been proposed to detect unusual score gains, collusion among candidates, aberrant wrong and right answer patterns, suspicious erasures and answer changes, and test retake violations, amongst others. Cizek (1999) supports the use of data forensics to detect cheating despite some limitations to the use of methods because “the conclusion that cheating has occurred is almost always probabilistic and requires inference” (p. 150). It is therefore common practice to rely on multiple sources of evidence and further investigation into statistical anomalies prior to taking action against any particular candidate or group of candidates. While there are still ongoing discussions regarding the legitimacy of enforcement of sanctions against candidates based on the results of data forensics, Maynes (2013) has suggested that these actions can be considered defensible provided that the methods employed implement proper error control and rely on accurate data, credible measurements, consistent procedures, scientific methods, probability statements, and well-reasoned findings. However, current data forensics range in their levels of effectiveness and sophistication (Fremer & Ferrara, 2013).

Although research in the field of data forensics has emerged and developed in recent years, more advanced methodologies to assess the impact of test fraud and item compromise “are in their infancy and little is known about how well the few methods work in practice” (Wollack & Fremer, 2013, p. 8). In Cizek’s (1999) review of statistical methods to identify students copying from one another, he noted that initial attempts to apply the Rasch person-fit measures to identify suspect candidates based on misfitting response patterns yielded little valuable information. However, more recent work in the application of IRT models—particularly Rasch—to generate statistics indicative of cheating behavior other than collusion have proven more successful. For example,

Maynes (2011) provided useful background on how IRT methods could be applied to compute score differences within a candidate's test responses through precise probability statements within single exam instances.

Item compromise calls into question the validity of candidates' test scores; therefore, researchers argue that analyses must compare the performance of candidates on both exposed and unexposed test content (Maynes, 2009; ATP, 2013; Carson, 2013). For example, Maynes' research (2011) purported the use of his score differential method to enable comparisons between new and old items, scored versus pretest items, and multiple choice and performance-based items to detect unusual score patterns within candidates that could indicate prior access to exam content. This chapter presents further methodology to enhance these data forensics techniques in terms of identifying aberrant performance by candidates in testing programs within largely exposed item pools. Namely, this chapter extends the existing research by Smith and Davis-Becker (2011) on how to detect candidates likely to have item preknowledge through the use of differential person functioning (DPF). It supports the utilization of DPF to identify candidates who likely gained prior access to exam content through illicit means by comparing performance on pretest items along with scored items. It then furthers the approach by following the DPF with subsequent differential item functioning (DIF) analyses to detect compromised items due to exposure. While this research does not address the current dearth of research "that has demonstrated the efficacy of DIF analysis for detecting group-based security breaches" (Maynes, 2013, p. 192), it does present evidence of how DPF can be used in conjunction with DIF to highlight individual candidate incidences of item compromise and reduce the impact of test fraud on item banks.

DATA

The data in this study were 3,280 administrations of data from a single test year of a continuously administered computer-based licensure program. The candidates were randomly administered one of two equated forms, with 1,636 candidates being administered Form 1 and 1,644 being administered Form 2. Each form consisted of 170 scored items and one of three different 10-item pretest sets, such that each candidate received 180 items. The overlap between forms (percentage/number of items shared between forms) was 51.8%, or 87 items.

METHODS

Differential Person Functioning (DPF)

As a first step, differential person functioning (DPF) was conducted using Winsteps (Linacre, 2009) to identify candidates likely to have had prior knowledge of exam content by comparing candidates' performance on scored and pretest items. DPF is a statistical analysis approach for comparing the performance of candidates on subsets of items while holding the item and person parameters constant, except for the person for whom DPF is being calculated. All other candidates and items are anchored at the Rasch measures from the initial calibration of all candidates and items. A candidate's ability measures are estimated on each subset of items, along with a calculation of the log-odds estimate of the difference between the two ability measures. Linacre (2009) notes that Winsteps calculates the person measure and standard error for the two item subsets for each candidate, and then a Mantel-Haenszel test is conducted. The

Mantel-Haenszel method is a log-odds estimator of DPF size and significance between the two item subsets, with the DPF contrast representing the log-odds estimate equivalent to a Mantel-Haenszel DPF size (Linacre, 2009). A *t*-test for significance in which the DPF contrast is divided by the joint standard error of the two DPF measures is then conducted, resulting in the probability of the likelihood of a particular combination of scores for each candidate.¹

For this analysis, the item subsets for the DPF are the 170 scored, operational items and the 10 pretest items administered to each candidate. This process enables flagging individual candidates with aberrant scores on the operational versus pretest items. Given the underlying assumptions of the Rasch model (i.e., sample independence), the precision of the ability estimates is not impacted by the comparative sample sizes of scored versus pretest items.

This methodology is based on the notion that candidates with prior knowledge of the item pool would likely have a high estimated ability on the scored items and a low estimated ability on the pretest items; this results in a low estimated probability of these two measures resulting for the same candidate. This presupposes that only the operational, scored items have been exposed and that the pretest items have not yet been subject to test fraud. If this condition is met, this DPF analysis provides evidence for a validity argument for or against candidates' exam scores by identifying candidates likely to have had prior content knowledge—intended or not.

The DPF analysis was conducted for all candidates to detect candidates with an unexpectedly low probability of the combination of their two ability estimates. Candidates were flagged as possibly suspect if they had more than a 1.0 logit difference between their respective ability measures on the scored items and pretest items (i.e., a DPF contrast greater than 1) and a probability of less than .05.

Though the focus of this DPF in this study was to detect candidates that likely had item preknowledge (significant positive DPF contrast), bidirectional analyses rather than directional analyses were run given that the test assumes that errors are random (so bidirectional analyses enhances the symmetry of the data) and to present a more conservative approach. Since the method is primarily designed to identify those candidates who performed differentially better on the scored versus pretest items, only the results for those candidates showing a significantly higher ability estimate on the scored items are presented.

Differential Item Functioning (DIF)

Differential item functioning (DIF) was subsequently conducted using Winsteps (Linacre, 2009) by comparing item difficulty for each item based on candidates' DPF results. The DIF procedure implemented in Winsteps is based on the same theoretical properties as the Mantel-Haenszel method (Linacre & Wright, 1987). The purpose of this step is to assess the extent to which candidates' prior knowledge of exam content impacted item performance. Of particular interest was the extent of item degradation that resulted from the unauthorized item exposure due to test fraud. Additionally, this step is critical for gathering information to drive exam maintenance, including identifying items in need of content refreshing or replacement as well as those appropriate to be utilized as anchor items.

DIF is a statistical approach for comparing item difficulty across subgroups while controlling for candidate ability and item difficulty, except for the item for which DIF is being calculated. All other items and candidates are anchored at the Rasch measure

from the initial calibration of all items and candidates. Item difficulties are calculated for each subgroup, along with a calculation of the log-odds estimate of the difference between the two difficulty measures. The analysis process implemented in Winsteps (Linacre, 2009) for DIF replicates that of DPF, such that the program calculates the item measure and standard error for the two subgroups for each item. The Mantel-Haenszel test is conducted, which is a log-odds estimator of DIF size and significance between the two subgroups, with the DIF contrast representing the log-odds estimate equivalent to a Mantel-Haenszel DIF size (Linacre, 2009). A *t*-test for significance in which the DIF contrast is divided by the joint standard error of the two DIF measures is then conducted, resulting in the probability of the likelihood of difference in difficulty for each candidate.²

For this DIF analysis, the subgroups of candidates were those flagged through the DPF versus those without flags, or candidates with likely prior item exposure versus those with response patterns not indicative of having item preknowledge. The intent of the DIF is to determine the extent to which items have been exposed by comparing the item difficulty measures for flagged versus nonflagged candidates. The assumption of this chapter is that noncompromised items would be of similar difficulty for both groups of candidates and that exposed items would favor flagged candidates. Again, given the underlying assumptions of sample invariance in the Rasch-based DIF model, the use of flagged candidates based on aberrant scores on scored versus pretest items should not affect model fit. The estimation of the parameters must be invariant across subsamples of candidates.

This DIF analysis was conducted on the entire 313 item pool (253 scored and 60 pretest items) to support exam maintenance in the context of a compromised item bank. Again, as with DPF, bidirectional DIF analyses were conducted with results being displayed for items with both significant positive and negative DIF; both sets of these results can help to inform future exam maintenance activities. Items were flagged if they had more than an absolute 1.0 logit difference between their respective difficulty measures for the flagged versus nonflagged candidates (DIF contrast greater than 1.0) and a probability of less than .05.

Selecting DPF and DIF Flagging Parameters

The number of candidates and items detected through the combined DPF and DIF approach varies depending on the flagging criteria selected, with the decision on flagging thresholds customizable to each respective exam's context and purpose. Factors influencing the selected criteria are two-fold, both psychometric and policy driven. Psychometrically, some factors to consider are the number of operational forms, candidate volumes, the proportion of various item types and status to the entire item bank, and the content refresh schedule. On the policy side, the testing program's capacity for follow-up investigation and enforcement, coordination with other security measures, and tolerance and resources for exam maintenance could impact the desired detection thresholds. For example, if defensible enforcement against suspect candidates is a primarily goal of the DPF analyses, conservative flagging criteria should be utilized to only identify candidates with the largest discrepancies between and lowest probabilities of their scores between the compromised and noncompromised content.

With regard to detecting compromised items, flagging criteria should be selected after considering the overall size of the item pool, tolerance for retiring and replacing items, ability to refresh content with new items, and budget and resources for

continued item development. For example, testing programs should take into account the following questions, among others: How many items are in the item pool, and what percent of those are exposed? To what extent has item degradation impacted item statistics and candidate results? Do sufficient unexposed items exist to retire and replace compromised items, or is new content necessary? What is the timeline of new item development, including review? Will the production and maintenance schedule allow for the piloting of new items?

The flagging criteria selected for this study were consistent across both the DPF and the DIF analyses. While this was considered to be appropriate for the purposes of this study, differential flagging criteria can certainly be utilized between the two analyses to arrive at an end result that best meets the needs of the testing program based on the determining factors described earlier. The DPF flagging criteria, which were more liberal than those suggested by Smith and Davis-Becker (2011), were considered appropriate for these analyses because the focus was identifying a cohort of candidates that likely had preknowledge of items for subsequent use as a subgroup in the DIF analyses. Had the objective been to provide probabilistic-based data to support enforcement cases and possible sanctions against particular candidates, it is likely that one would want to be more conservative with the flagging criteria (e.g., the criteria utilized by Smith and Davis-Becker [2011] were a DPF contrast greater than 3 and a probability less than .0001).³ The DIF flagging criteria were also considered appropriate for these analyses as the purpose of the DIF was to detect compromised items for retirement (significant negative DIF contrast) as well as well-functioning items that could be presumed secure and appropriate to use for equating, anchoring, and/or scored items on future exam iterations (significant positive DIF contrast). In circumstances with more limited support for wide-scale item development, replacement, and exam maintenance, more conservative DIF contrast and probability flagging criteria would likely be better to reduce the pool of compromised items identified through the analyses.

RESULTS

Differential Person Functioning (DPF)

Of the 3,280 candidates, 56 candidates (1.7%) were flagged for DPF based on the established criteria. The critical probability to identify candidates with DPF was set to $p \leq 0.05$, with a DPF contrast (difference between the estimated Rasch ability measures) ≥ 1.0 . In this analysis, a positive DPF contrast suggests that the candidate scored significantly better on the scored items than on the pretest items. Table 7.1 contains the overall DPF results.

Figure 7.1 displays the contrasts in the DPF ability measures for each of the candidates and indicates those candidates with the most differential results by scored versus pretest Rasch measures. As seen in Figure 7.1, 56 candidates scored unexpectedly well

Table 7.1 DPF Results by Item Type

Number of Responses	Items Per Candidate		% Flagged for Suspected Preknowledge	Number of Flagged Candidates			
	Scored	Pretest		Contrast ≥ 1	Contrast ≥ 2	Contrast ≥ 3	
				3,280	170	10	1.71%

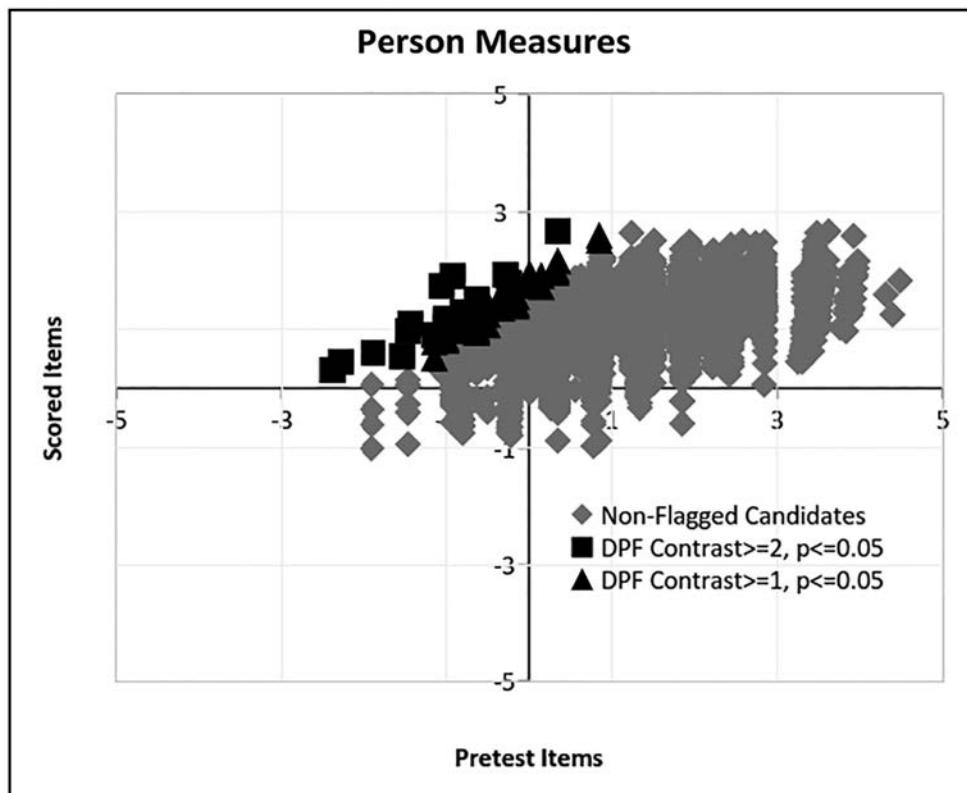


Figure 7.1 Differential Person Functioning by Scored vs. Pretest Items

on the scored items as compared to the pretest items (e.g., DPF measures of 1.91 and -0.92, respectively, on the scored and pretest items, DPF contrast of 2.83). These candidates were therefore considered most likely to have had item preknowledge and were flagged as suspect for having access to exam content prior to exam administration.

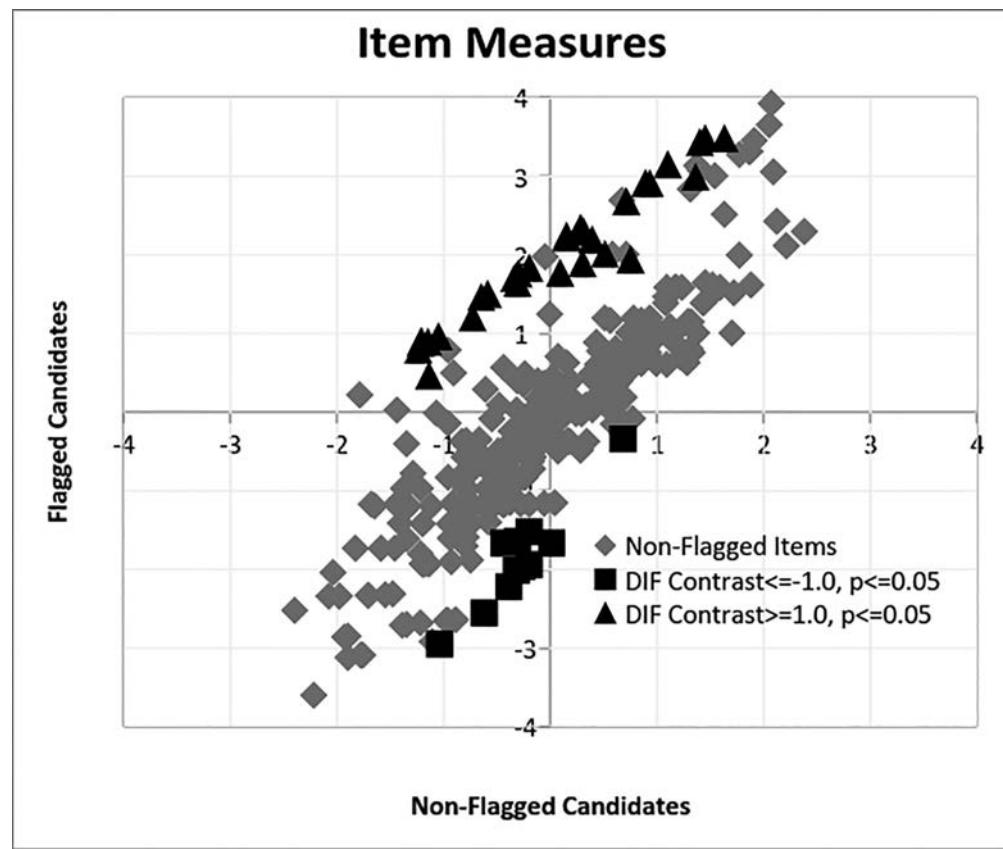
Differential Item Functioning (DIF)

Of the 313 total items (60 pretest items, 253 scored items), 41 items (13.1%) displayed significant DIF based on the set parameters. The critical probability to identify items with DIF was set to $p \leq .05$. The DIF contrast (i.e., the absolute difference between the estimated Rasch item difficulties) was $\geq |1.0|$. In these analyses, a positive DIF contrast suggests that the item is more difficult for the flagged candidates; while a negative DIF contrast suggests the item is less difficult for the flagged candidates. The items of most interest were therefore those easier for the flagged candidates (significant negative DIF) as these indicate items with potential compromise. The items with significant positive DIF were also identified to determine any items that were harder for candidates with presumed prior knowledge of the exam content, as these indicate items that could be considered most secure and viable for use as anchor items or continued use as either pretest or scored items. Table 7.2 contains the overall DIF results by item type.

Figure 7.2 displays the DIF results by candidate flagging status. As seen in Figure 7.2, the DIF results showed ten scored items that were statistically significantly easier for

Table 7.2 DIF Results by Item Type

Number of Unique Items	Item Status		% Items Flagged for Suspected Compromise		% Items Flagged as Presumed Secure		Number of Flagged Items			
			Scored	Pretest	Scored	Pretest	Scored	Pretest	Contrast ≤ -1.0	Contrast ≥ 1.0
	Scored	Pretest	Scored	Pretest	Scored	Pretest	10	31		
313	253	60	3.95%	0.00%	1.67%	50.00%	10	31		

**Figure 7.2** Differential Item Functioning by Candidate Flagging Status

flagged candidates, such that flagged candidates performed better on these items than nonflagged candidates (e.g., DIF measures of -2.55 and -0.62, respectively, for flagged and nonflagged candidates, DIF contrast of -1.93). These items could therefore be considered as the most grossly compromised and be marked for retirement and replacement. In contrast, 31 items (30 pretest and one scored) displayed DIF in the opposite direction; these items were significantly harder for flagged candidates (e.g., DIF measures of 1.89 and 0.30, respectively for flagged and nonflagged candidates, DIF contrast of 1.58). These items could therefore be considered as the most secure, least subject to piracy, the best items for anchoring, equating, use as common scored items, and/or continued use as pretest items on future iterations of the exam.

DISCUSSION

To further demonstrate the feasibility and benefit of this approach, DPF analyses were conducted on seven large-scale certification exams across multiple organizations. DIF analyses were also run on three of the exams to extend the interpretability of the DPF flagging results into exam maintenance and show a range of DIF results from which a testing program can determine best next steps for particular items. This comparison of scored items to unscored items has been shown to be useful in identifying candidates likely entering the test with prior exam knowledge in past research (Smith & Davis-Becker, 2011; O'Leary & Smith, 2013). These additional analyses further demonstrate how the combination of these methodologies provides several practical options for testing programs trying to maintain the validity of their candidate decisions in the context of grossly compromised item banks.

Each of the seven exams had presumed item-exposure issues, with evidence of item compromise displayed through increases in moving average scores over time and/or high incidence of candidates with high exam scores in low time on mid-administration form- and item-level analyses. For each of these exams, candidates were randomly administered one of multiple preequated, parallel forms, each consisting of a set of scored items and a set of unscored items. In all cases, these unscored items represented newly written pilot items and/or revised existing items previously flagged for an option or key issue.

For comparison purposes, the flagging parameters for the DPF and DIF analyses were both consistent with the computer-based licensure program study and held constant across all exams. As such, candidates were flagged as possibly suspect if they had more than a 1.0 logit difference between their respective ability measures on the scored and unscored items and $p \leq .05$. Likewise, items were flagged as potentially compromised and/or secure if they had more than a $|1.0|$ logit difference between item difficulty for flagged and nonflagged candidates and $p \leq .05$. Table 7.3 presents the DPF results by exam, including the total number of candidates, number and status of items administered to each candidate, and proportion of candidates flagged.

As shown in Table 7.3, candidates were flagged for anomalous differences between their scored and unscored exam scores on all seven large-scale certification exams. The proportion of suspect candidates detected ranged from less than 1% to more than 10%. While these organizations were approaching exam security issues (prevention, detection, and enforcement) through varying approaches with differing levels of rigor, these

Table 7.3 DPF Results by Large-Scale Certification Exam

Exam	Number of Candidates	Items Per Candidate		% Flagged for Suspected Preknowledge	Number of Flagged Candidates	
		Scored	Unscored		Contrast ≥ 1	Contrast ≥ 2
Exam 1	3,946	61	14	0.41%	16	0
Exam 2	1,547	54	20	3.56%	52	3
Exam 3	3,812	42	18	4.20%	147	13
Exam 4	507	80	20	4.34%	22	0
Exam 5	341	39	26	4.69%	9	7
Exam 6	9,974	45	15	8.11%	664	145
Exam 7	2,155	41	19	11.18%	119	122

Table 7.4 DIF Results by Large-Scale Certification Exam

Exam	Number of Unique Items	Item Status		% Items Flagged for Suspected Compromise		% Pretest Items Flagged for Security		Number of Flagged Items	
		Scored	Unscored	Scored	Unscored	Scored	Unscored	<= -1.0	>= 1.0
Exam 1	268	144	124	0.69%	0.00%	0.69%	3.23%	1	5
Exam 3	222	73	149	41.10%	0.00%	0.00%	18.79%	30	28
Exam 7	251	74	177	67.57%	0.00%	0.56%	50.28%	50	90

aggregate results indicate the widespread utility of utilizing DPF to identify anomalous candidate behavior apart from other security strategies.

Table 7.4 presents the DIF results by exam, including the total number and status of items in the item pool, proportion of items flagged for suspected compromise (negative DIF contrast), proportion of items flagged as presumed secure (positive DIF contrast), and the total numbers of flagged items.

As shown in Table 7.4, at least a few items were flagged for anomalous differences in item difficulty for suspect and nonsuspect candidates on each of the three large-scale certification exams for which DIF was conducted. The proportion of compromised content ranged from less than 1% to more than 67% of scored items. The proportion of content considered to be most secure and least subject piracy at that point in time ranged from approximately 3% to over 50% of unscored items. These analyses show the benefit of following the DPF analysis used to detect suspect candidates with DIF analyses to identify problematic and well-functioning items, and the varying results across different exams. The testing programs for these exams could employ various exam maintenance activities based upon these results, ranging from removal and replacement of the compromised items, development of new items in impacted content areas, increased use of the secure items as anchor items, equating items, and/or scored items, or reduction in time between content refreshes and forms reassemblies.

CONCLUSIONS

The rampant test fraud in certification testing programs has led to the widespread unauthorized exposure of exam forms and perpetual item compromise. As shown in this chapter, the combination of DPF and DIF can strengthen the security monitoring of a testing program with known exposure issues. The tight timeframes for piracy (within days or weeks) within certification testing programs gravely reduce the benefit of investing the budget and resources into developing an entirely new item bank for exam with egregious exposure issues; that content would likely be stolen and exposed too quickly for the testing program to benefit from its efforts. Instead, realistic approaches to exam and item maintenance are needed to reduce threats to the validity of score interpretation and use in continually administered exams with exposure issues by highlighting specific compromised content.

This chapter presented how the use of DPF in conjunction with DIF can minimize the tangible and intangible costs of test fraud, in addition to maximizing the measurement integrity and validity associated with exams with this known security issue. Additionally, these procedures could deter future test fraud by enhancing the data

forensics associated with a testing program. Wollack and Fremer (2013) note that well-articulated procedures are “a very effective way to communicate to candidates that cheaters leave behind irregular patterns of responses, and that even if they are sufficiently clever to successfully cheat on the exam, they will be unearthed by sophisticated statistical procedures being run in the background” (p. 11). It should be noted that these techniques rely on the hypothesis that only operational items have been exposed, and that pretest, pilot items have not been subject to test fraud. This assumption should be investigated and accepted by a given testing program prior to implementing these security analyses.

With that caveat, the combined use of DPF and DIF enhances a test program’s ability to maintain credible exams within a context of prevalent test fraud by (1) detecting suspect candidates by their aberrant exam scores by item status and (2) identifying compromised and secure items based on bias towards or against candidates flagged for item preknowledge. Because DPF incorporates probabilistic-statements regarding discrepancies in individual candidates’ comparative scores within a single testing instance on a respected measurement model (Rasch), the results do align with Maynes’s (2013) requirements for enforceable security analyses. These analyses could therefore lead to defensible sanctions against candidates when combined with other evidence. In these situations, candidates’ scores could be invalidated or candidates could be banned from testing for a set period of time, among a myriad of other actions. Likewise, the DIF analysis as described presents probabilistic statements about apparent bias in particular items favoring those candidates with item preknowledge. Therefore, the extent of item degradation can be controlled by regularly checking the extent of exposure for particular items in compromised item banks. Once compromise is detected for particular items, these items can either be replaced immediately with pilot items or be monitored for changes in their item statistics and replaced as necessary and feasible, given the status of item development and availability of new items.

Research still needs to be conducted to confirm these results and explore extensions of these methods, such as scale stability over time. This chapter presents how these methods can be used to (1) detect when security breaches have occurred; (2) determine the extent of item compromise; (3) build cases against suspect candidates; (4) collaborate with other evidence to support the enforcement of sanctions against candidates; (5) highlight specific items with compromised content, and (6) evaluate appropriate next steps for particular items and entire item banks while discussing the relevant psychometric and policy issues for each of these areas. The application of these dual analysis efforts will help to preserve the validity of candidate decisions and the reputation of testing programs operating in an environment of grossly exposed exam content. This chapter also contributes to the growing literature on data forensic techniques to gauge the impact of test fraud on testing programs via statistical and psychometric methods.

NOTES

1. Refer to www.winsteps.com/winman/difconcepts.htm for additional details of and formulas for the calculation of DPF in Winsteps (Linacre, 2009).
2. Refer to www.winsteps.com/winman/difconcepts.htm for additional details of and the formulas for the calculation of DIF in Winsteps (Linacre, 2009).
3. Also see this research for practical guidance on decision consistency (as well as Type I and Type II error rates) which can be expected for a variety of pretest item sample sizes that could be useful to selecting flagging criteria.

REFERENCES

- Association of Test Publishers (ATP) Security Committee (2013, January). Assessment security options: Considerations by delivery channel and assessment model. Retrieved April 11, 2013 from www.testpublishers.org/assets/assessment_security_options_considerations_by_delivery_channel_and_assessment_model_1-23-13.pdf.
- Carson, J. D. (2013). Certification/licensure testing case studies. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 261–283). New York, NY: Routledge.
- Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it and prevent it*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Foster, D. (2013). Security issues in technology-based testing. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 39–83). New York, NY: Routledge.
- Fremer, J. J. & Ferrara, S. (2013). Security in large-scale paper and pencil testing. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 17–37). New York, NY: Routledge.
- Han, N. (2003). Using moving averages to assess test and item security in computer-based testing. *Center for Educational Assessment Research Report No. 468*. Amherst, MA: University of Massachusetts, School of Education.
- Han, N. & Hambleton, R. K. (2004, April). *Detecting exposed test items in computer-based testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Han, N. & Hambleton, R. K. (2008). Detecting exposed items in computer-based testing. In C. L. Wild & R. Ramaswamy (Eds.), *Improving testing: Applying process tools and techniques to assure quality* (pp. 423–448). New York, NY: Lawrence Erlbaum Associates.
- Impara, J. C. & Foster, D. (2006). Item development strategies to minimize test fraud. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 91–114). Mahwah, NJ: Lawrence Erlbaum Associates.
- Linacre, J. M. (2009). WINSTEPS® Rasch measurement computer program. Beaverton, OR: Winsteps.com.
- Linacre, J. M. & Wright, B. D. (1987). Item bias: Mantel-Haenszel and the Rasch model. *Memorandum No. 39 MESA Psychometric Laboratory*. Chicago, IL: University of Chicago, Department of Education.
- Lu, Y. & Hambleton, R. K. (2003). Statistics for detecting disclosed item in a CAT environment. *Center for Educational Assessment Research Report No. 498*. Amherst, MA: University of Massachusetts, School of Education.
- Maynes, D. (2009). Caveon speaks out on exam security: The last five years. Retrieved April 10, 2013 from www.caveon.com/articles/_Exam_Security.pdf.
- Maynes, D. D. (2011, April). *A method for measuring performance inconsistency by using score differences*. Chicago, IL: Paper presented at the annual conference of the Society for Industrial and Organizational Psychology.
- Maynes, D. (2013). Educator cheating and the statistical detection of group-based test security threats. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 173–199). New York, NY: Routledge.
- O'Leary, L. S. & Smith, R. W. (2013, April). *Extending differential person and item functioning to aid in maintenance of exposed exams*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Smith, R. W. (2004, April). *The impact of braindump sites on item exposure and item parameter drift*. Paper presented at annual meeting of the American Educational Research Association, San Diego, CA.
- Smith, R. W. & Davis-Becker, S. (2011, April). *Detecting suspect candidates: An application of differential person functioning analysis*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Veerkamp, W. J. J. & Glas, C. A. W. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, 25, 373–389.
- Wollack, J. A. & Fremer, J. J. (2013). Introduction: The test security threat. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 1–13). New York, NY: Routledge.
- Yang, Y., Ferdous, A., Chin, T. Y. (2007, April). *Exposed items detection in personnel selection assessment: An exploration of new item statistic*. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

8

IDENTIFICATION OF ITEM PREKNOWLEDGE BY THE METHODS OF INFORMATION THEORY AND COMBINATORIAL OPTIMIZATION

Dmitry Belov

INTRODUCTION

Item preknowledge occurs when some examinees (called *aberrant examinees*) have had prior access to a subset of items (called *compromised items*) administered on a live test form. As a result, aberrant examinees perform better on compromised items as compared to uncompromised items. When the number of aberrant examinees is large, the corresponding testing program and its stakeholders (universities, companies, government organizations, etc.) are negatively affected because they are given invalid scores for aberrant examinees.

In general, item preknowledge is hard to detect due to multiple unknowns involved: unknown subgroups of examinees (at unknown schools or test centers) accessing unknown compromised subsets of items prior to taking the test. Recently, multiple statistical methods were developed to detect compromised items by exploiting certain test characteristics specific to each testing program (Choe, 2014; Obregon, 2013; O'Leary & Smith, 2013; van der Linden & Guo, 2008). However, the detected item subset naturally has some uncertainty due to false positives and false negatives. The uncertainty increases when different subgroups of aberrant examinees had access to different subsets of items; thus, compromised items for one subgroup are uncompromised for another subgroup and vice versa (Belov, 2015). This chapter will describe an algorithm to detect all three unknowns, leading to better detection of aberrant examinees and better identification of the specific items to which each had prior access. The advantages and limitations of the algorithm are demonstrated on simulated and real data. This algorithm extends the 3D Algorithm by Belov (2014).

PROBLEM STATEMENT

Each examinee has a profile, which includes variables potentially useful for test security purposes: test center where the test is taken, former high school, former undergraduate

college, attended test-prep center, or current group in a social network. Each variable can be used to partition examinees into disjoint groups that are homogeneous with respect to their potential for item sharing, thereby allowing separate investigations to be conducted for each group. For example, the following variables partition examinees by geographic location where examinees take a test: room, college, state, region, country. Variables related to geographic location are most common. However, as Belov (2013) pointed out, other profile variables from above could potentially help to detect aberrant examinees, even if they take exam at different geographic locations.

Consider a profile variable (or a combination of profiles variables) partitioning all examinees into disjoint *groups*. A group without any aberrant examinee (this group has only nonaberrant examinees and, therefore, it is unaffected by the item preknowledge) is called an *unaffected group*. A group with nonaberrant and aberrant examinees is called an *affected group*. Within each affected group its aberrant examinees can be represented by disjoint *aberrant subgroups*, where each aberrant subgroup has preknowledge of a unique *compromised subset* of items (see Figure 8.1). Note that while aberrant subgroups are disjoint, compromised subsets are not, suggesting that multiple aberrant subgroups may have preknowledge of the same items (Figure 8.1). The above terms are chosen to avoid confusion in the text: *group* and *subgroup* refer only to examinees; *set* and *subset* refer only to items (a reader should follow this terminology to understand this chapter).

Two assumptions are made in this chapter. The first assumption is that a test consists of a fixed number of highly correlated operational sections (adaptive or not), each measuring the same underlying construct. For example, the Law School Admission Test (LSAT) consists of four operational sections (with approximately 25 items each): one with analytical reasoning (AR) items, two with logical reasoning (LR) items, and one with reading comprehension (RC) items. Items from each operational section are pretested in previous administrations. Therefore, it is possible that some of the items presented in operational sections are compromised. In many other testing programs, operational items may also have been used operationally before, thereby increasing their exposure even more.

The second assumption is that each affected group cannot have more than one aberrant subgroup per operational section (see Figure 8.1). Considering how small a group can be (e.g., class) and how small the operational section can be, this assumption appears to be realistic. For example, in case of the LSAT, an affected group can have four aberrant subgroups, where aberrant subgroups 1, 2, 3, and 4 have preknowledge of items from different operational sections (e.g., from sections 1, 2, 3, and 4, respectively); at the same time, an affected group cannot have two aberrant subgroups, where aberrant subgroups 1, 2 have preknowledge of items from the same operational section (e.g., from section 3).

The item preknowledge detection problem addressed in this chapter is now stated as follows:

How does one detect *affected groups*, their *aberrant subgroups*, and corresponding *compromised subsets*?

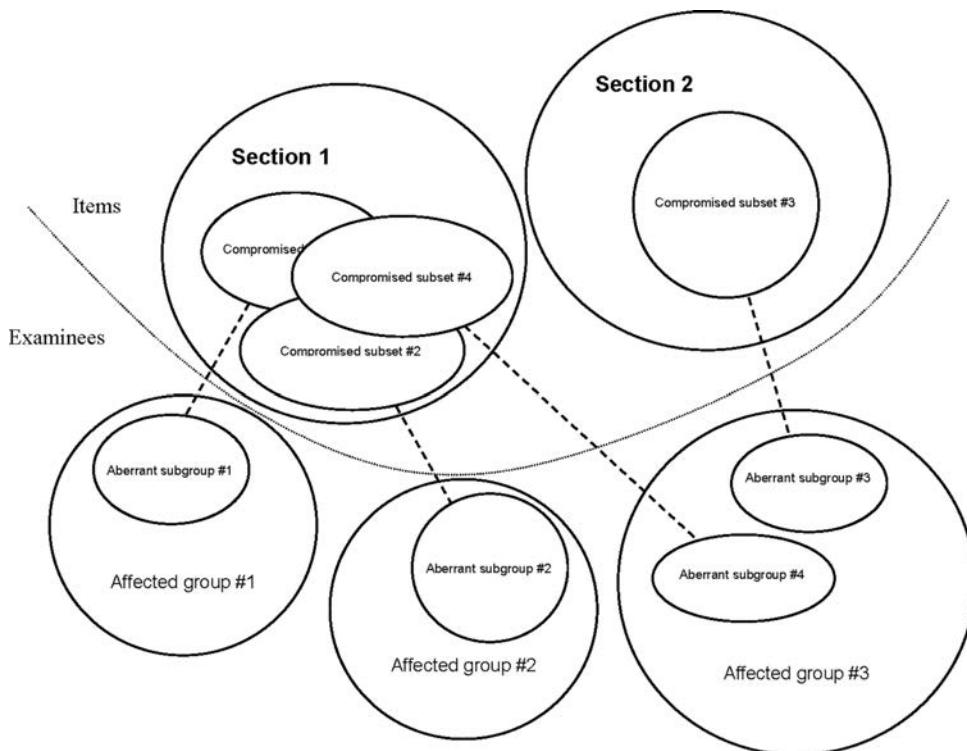


Figure 8.1 Terms of Item Preknowledge and Their Interaction for an Example of Test Consisting of Two Operational Sections

TWO ITEM PREKNOWLEDGE DETECTION STATISTICS

Before tackling the item preknowledge detection problem, let us consider two statistics for detecting item preknowledge. Throughout the chapter the following notation is used:

- Lowercase letters a, b, c, \dots denote scalars.
- Lowercase Greek letters $\alpha, \beta, \gamma, \dots$ denote random variables; an estimator of a random variable is shown using a caret (e.g., an estimate of θ is denoted as $\hat{\theta}$).
- Capital letters A, B, C, \dots denote sets (or subsets) of items and groups (or subgroups) of examinees.
- Capital Greek letters $\Omega, \Psi, \Theta, \dots$ denote collections of subsets.
- Bold capital letters $\mathbf{A}, \mathbf{B}, \mathbf{C}, \dots$ denote functions.

An examinee is defined by two random variables: an unobservable latent trait (ability) θ and an observable response $\chi_i \in \{0, 1\}$ to item i . Consider an arbitrary subset of items I administered to the examinee, where I may vary among examinees (e.g., CAT) or be fixed (e.g., P&P). Then, Bayes' theorem is applied to compute the discrete posterior distribution of θ with a uniform prior distribution:

$$F_I(y) = \frac{\prod_{i \in I} P_i(\chi_i | y)}{\sum_{z \in Y} \prod_{i \in I} P_i(\chi_i | z)}, \quad y \in Y, \quad (1)$$

where $\mathbf{F}_I(y)$ is the probability of $\theta = y$, $\mathbf{P}_i(\chi_i|y)$ is the probability of response χ_i to item i conditioned on $\theta = y$, and set Y contains ability levels (this chapter used $Y = \{-5, -4.8, \dots, 5\}$). Hence, the expected a posteriori (EAP) estimator of θ is computed as follows:

$$\hat{\theta}_I = \sum_{y \in Y} y \mathbf{F}_I(y). \quad (2)$$

From an aberrant examinee standpoint, the administered test is partitioned into two disjoint subsets: first subset with compromised items and second subset with uncompromised items. Naturally, this examinee will perform better on the first subset in contrast to the second subset. Therefore, a statistic to detect aberrant examinees may be based on some measurement of the performance change between first and second subsets: the larger the change, the higher the probability that the given examinee is aberrant. As it was emphasized in the Introduction, the aberrant examinees and compromised items are unknown and to be detected. For the sake of simplicity, let us assume in this section of the chapter that there is only one compromised subset, S , and it is known to us.

Consider an examinee taking a test T (where T may vary among examinees), which can be partitioned into two disjoint subsets $C = T \cap S$ (compromised items, intersection of the test and the compromised subset) and $U = T \setminus S$ (uncompromised items, test items without items from the compromised subset). Then the following characteristics can be computed for C , U , and T : posterior distributions of ability, \mathbf{F}_C , \mathbf{F}_U , \mathbf{F}_T and ability estimates $\hat{\theta}_C$, $\hat{\theta}_U$, $\hat{\theta}_T$. These characteristics will be used to describe two statistics for detecting aberrant examinees. Each statistic is computed such that aberrant examinees should be located at the right tail of the corresponding null distribution.

Statistic lz

This statistic is computed as follows:

$$\begin{aligned} & -\frac{\lambda - e_\lambda}{v_\lambda} \\ & \lambda = \sum_{i \in T} \chi_i \ln \mathbf{P}_i(1|\hat{\theta}_T) + (1 - \chi_i) \ln \mathbf{P}_i(0|\hat{\theta}_T) \\ & e_\lambda = \sum_{i \in T} \mathbf{P}_i(1|\hat{\theta}_T) \ln \mathbf{P}_i(1|\hat{\theta}_T) + \mathbf{P}_i(0|\hat{\theta}_T) \ln \mathbf{P}_i(0|\hat{\theta}_T) \\ & v_\lambda = \sum_{i \in T} \mathbf{P}_i(1|\hat{\theta}_T) \ln \mathbf{P}_i(1|\hat{\theta}_T) \left(\ln \frac{\mathbf{P}_i(1|\hat{\theta}_T)}{\mathbf{P}_i(0|\hat{\theta}_T)} \right)^2 \\ & \mathbf{P}_i(0|\hat{\theta}_T) = 1 - \mathbf{P}_i(1|\hat{\theta}_T). \end{aligned} \quad (3)$$

Drasgow, Levine, and Williams (1985) introduced lz based on the true value θ ; however, this value is unknown in practice, and an estimate $\hat{\theta}_T$ [see Equation (2)] is commonly used instead. Statistic (3) does not rely on subset S and is often used as a baseline in computational studies on detecting item preknowledge (Belov, 2014, 2015; Levine & Drasgow, 1988; Shu, Henson, & Leucht, 2013). The properties of lz are well explored (Armstrong, Stoumbos, Kung, & Shi, 2007). The asymptotic distribution of lz converges to $N(0, 1)$ only when $\theta = \hat{\theta}_T$, which scarcely happens in practice.

Divergence Statistic

The Kullback–Leibler divergence (KLD) is a widely used measure of similarity between two distributions. The higher its value the lower the similarity and vice versa, where the lowest value (which is zero) is reached when the two distributions are identical. Therefore, one can apply KLD to measure difference between F_U and F_C (see above), which corresponds to a performance change between compromised and uncompromised items. The use of KLD for detecting aberrant responding was first proposed by Belov, Pashley, Lewis, and Armstrong (2007).

Formally, KLD for discrete distributions F_U and F_C is computed as follows (Kullback & Leibler, 1951):

$$D(F_U \parallel F_C) = \sum_{y \in Y} F_U(y) \ln \frac{F_U(y)}{F_C(y)}. \quad (5)$$

According to the definition of KLD, the larger the divergence $D(F_U \parallel F_C)$, the higher the dissimilarity between F_U and F_C . The value of $D(F_U \parallel F_C)$ is always nonnegative and equals zero only if the two distributions are identical. KLD is asymmetric; that is, in general, $D(F_U \parallel F_C) \neq D(F_C \parallel F_U)$.

One issue with application of KLD for detecting item preknowledge is illustrated in Figure 8.2, where posterior distributions of ability are shown for two examinees: a nonaberrant examinee performs on subset U better than on subset C , see Figure 8.2(a); an aberrant examinee performs on subset C better than on subset U , see Figure 8.2(b). In both cases KLD has the same value, 2, which, potentially, makes both examinees detected or undetected simultaneously, thus, increasing Type I and Type II errors. Fortunately, this issue is easy to fix by assigning KLD to zero when an examinee does not perform on compromised items better than on uncompromised items, as follows:

$$D^*(F_U \parallel F_C) = \begin{cases} \sum_{y \in Y} F_U(y) \ln \frac{F_U(y)}{F_C(y)}, & \hat{\theta}_U \leq \hat{\theta}_C \\ 0, & \hat{\theta}_U > \hat{\theta}_C \end{cases} \quad (6)$$

Thus, $D^*(F_U \parallel F_C) = 0$ for the nonaberrant case from Figure 8.2(a) and $D^*(F_U \parallel F_C) = D(F_U \parallel F_C) = 2$ for the aberrant case from Figure 8.2(b). The statistic defined by Equation (6) will be called here a *divergence statistic*.

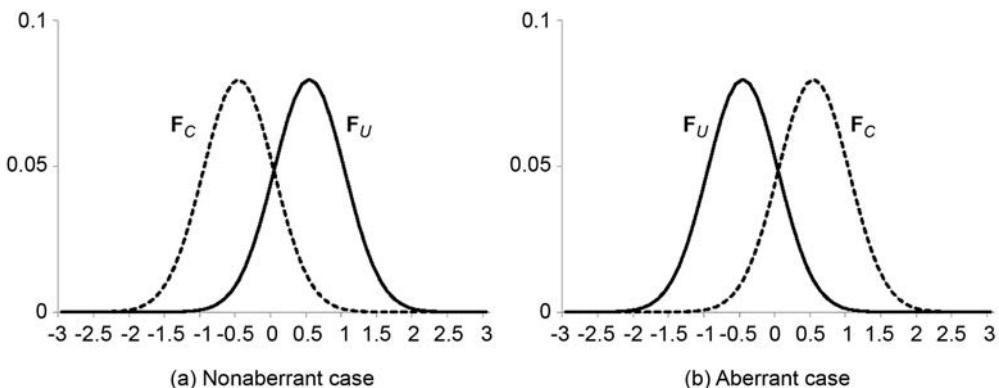


Figure 8.2 Posterior Distributions of Ability for Nonaberrant Examinee (a) and Aberrant Examinee (b)

DIVERGENCE ALGORITHM

An algorithm to solve the item preknowledge detection problem is introduced here. It is a modification of the 3D Algorithm by Belov (2014). When searching for compromised subset only responses of examinees from a suspicious subgroup (see below) are taken into account, which allows detecting smaller aberrant subgroups and smaller compromised subsets in contrast with the 3D Algorithm. Also, this modification can detect multiple aberrant subgroups per affected group (see Figure 8.1). This is achieved by analyzing each operational section O_1, O_2, \dots, O_n separately, where n is number of operational sections. Thus, each examinee can be detected multiple times. Therefore, to prevent overdetection of examinees, the Bonferroni correction (Abdi, 2007) is applied for detecting aberrant examinees, i.e., the significance level for the divergence statistic is divided by n .

For each operational section $O_i, i = 1, 2, \dots, n$ the following steps will be performed. Let us generate a collection of random subsets $\Omega = \{S_1, S_2, \dots, S_m\}, S_k \subset O_i, k = 1, 2, \dots, m$ and compute divergence statistic $d_{j,S} = \mathbf{D}^*(\mathbf{F}_{T_j \setminus S} || \mathbf{F}_{T_j \cap S})$ for each examinee j and for each $S \in \Omega$, where T_j is a test administered to examinee j . If an examinee j is aberrant and the corresponding compromised subset belongs to O_i then distribution of the divergence statistic over Ω will be positively skewed and will have an unusually high expectation $E_\Omega[d_{j,S}]$ because multiple subsets from the Ω will intersect the compromised subset. The expectation $E_\Omega[d_{j,S}]$ is used to detect affected groups and *suspicious subgroups* (subgroups of examinees that may intersect the corresponding aberrant subgroups) as follows: a group is affected if at least one of its examinees j has $E_\Omega[d_{j,S}] > C(\alpha_1)$, then each examinee j from the affected group with $E_\Omega[d_{j,S}] > C(\alpha_2)$ is included into corresponding suspicious subgroup J . Critical values $C(\alpha_1)$ and $C(\alpha_2)$ are computed from simulated examinees drawn from $N(0, 1)$ population, where $\alpha_1 < \alpha_2$. In each affected group, for a variable $S \subset O_i$ let us compute the expectation of $d_{j,S}$ over examinees from the suspicious subgroup J , denoted as $E_J[d_{j,S}]$. If multiple examinees in the suspicious subgroup J are aberrant then, due to definition of the divergence statistic, the $E_J[d_{j,S}]$ should be maximized at the corresponding compromised subset. Thus, one can use $E_J[d_{j,S}]$ as an objective function for a combinatorial search for the corresponding compromised subset S^* (this search is described in the next section). Finally, for each affected group its examinees with $d_{j,S^*} > C(\alpha_3)$ are included into aberrant subgroup, where S^* is found compromised subset and the critical value $C(\alpha_3)$ is computed from simulated examinees drawn from $N(0, 1)$ population. The following represents formal steps of the algorithm:

Divergence Algorithm

For each operational section $O_i, i = 1, 2, \dots, n$ run the following:

Step 1: Detect affected groups and suspicious subgroups via statistic $E_\Omega[d_{j,S}]$.

Step 2: For each affected group, detect compromised subset $S^* \subset O_i$ by solving $\max_{S \subset O_i} E_J[d_{j,S}]$, where S is optimization variable and J is detected suspicious subgroup.

Step 3: For each affected group and compromised subset $S^* \subset O_i$ detect aberrant subgroup via statistic d_{j,S^*} .

The success of the Divergence Algorithm is hinged on how large the intersection (meaning percentage of aberrant examinees in the suspicious subgroup) of the suspicious subgroup with the actual aberrant subgroup. If this intersection is not large, then the search of the compromised subset may fail. Intuitively, this should not happen because of Step 1, where multiple random subsets are generated resulting in a large coverage of the actual compromised subset. Such coverage should guarantee a correct detection of the affected group and a large intersection of the suspicious subgroup with the actual aberrant subgroup.

SEARCH FOR THE COMPROMISED SUBSET VIA SIMULATED ANNEALING

This section will describe a combinatorial search for the compromised subset given a fixed suspicious subgroup (see Step 2 of the Divergence Algorithm). The search is based on a special iterative process called simulated annealing. Simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983) is a heuristic for finding an optimal solution to a given unconstrained optimization problem. The name comes from annealing in metallurgy, a technique involving heating and controlled cooling of a material to reduce its defects. The convergence of simulated annealing can be analyzed by its reduction to a Markov chain (Bertsimas & Tsitsiklis, 1993).

For each suspicious subgroup J , the simulated annealing starts with an initial solution that provides the largest value of $E_J[d_{j,S}]$, where $S \in \Omega$ (see above). The initial solution is assigned to the best solution S^* (detected compromised subset) and a current solution S_0 (corresponding to a current iteration of the optimization process). Multiple iterations are performed to improve S^* by exploring on each iteration the following randomly chosen modifications of S_0 :

Modification:	1	2	3	(7)
Probability:	P(1)	P(2)	P(3),	

where modification 1 adds a random item from $O_i \setminus S_0$ to S_0 (O_i is current operational section, see above); modification 2 swaps a random item from $O_i \setminus S_0$ with a random item from S_0 ; and modification 3 removes a random item from S_0 .

At each iteration, if a modification randomly chosen according to (7) results in an improvement of the best solution S^* , then this modification is accepted for both S^* and S_0 ; otherwise, this modification is accepted only for S_0 with a probability, which is gradually decreasing, and this nonlinear decrease is controlled by a parameter t called temperature. Accepting nonoptimal modifications prevents getting stuck in a local maximum.

Simulated Annealing Process to Search for Compromised Subset

Step 1: Set the best solution $S^* = \arg \max_{S \in \Omega} E_J[d_{j,S}]$, the current solution $S_0 = S^*$, and the temperature $t = t_0$.

Step 2: Set subset $S = S_0$, then simulate random variable $\delta \in \{1, 2, 3\}$ according to the discrete distribution (7) and modify S , respectively.

Step 3: If $E_J[d_{j,S}] > E_J[d_{j,S^*}]$, indicating that an improvement to the best solution has been found, then set $S_0 = S$ and $S^* = S$ (update the best solution) and go to Step 5; otherwise continue to Step 4.

- Step 4: Simulate a uniformly distributed $\gamma \in [0, 1]$. If $\gamma < \exp\left(\left[\mathbf{E}_J[d_{j,S}] - \mathbf{E}_J[d_{j,S_0}]\right]/t\right)$ (the probability of accepting a modification to the current solution, S_0 , that did not improve the best solution, S^*) then set $S_0 = S$ (update the current solution).
 Step 5: If $t > t_1$ then $t = t \times d$ and go to Step 2 (perform more iterations to improve the best solution); otherwise stop (S^* is detected compromised subset).

SIMULATED DATA EXPERIMENTS

Two approaches to detect item preknowledge were compared. The first approach was to compute the I_z statistic (Drasgow, Levine, & Williams, 1985), which was used as a baseline in all studies. The second approach was the Divergence Algorithm. Both approaches were implemented in C++ by the author. All critical values were computed from simulated data, where 10,000 nonaberrant examinees were drawn from $N(0, 1)$ distribution.

Simulation Design

The disclosed operational LSAT form was used. It had four operational sections, each with approximately 25 items: one with analytical reasoning (AR) items, two with logical reasoning (LR) items, and one with reading comprehension (RC) items.

Ten test centers were simulated, each with 100 examinees. Nonaberrant examinees were drawn from $N(0, 1)$ population. Aberrant examinees were drawn from $U(-3, 0)$ population. The probability of examinees from aberrant subgroups correctly answering items from corresponding compromised subsets was 1 (see Figure 8.1); otherwise, the probability was modeled by a three-parameter logistic model (Lord, 1980). The number of affected test centers, aberrant subgroups, and examinees within subgroup varied across experiments. For each aberrant subgroup, its compromised subset was formed in two steps: first, an operational section was randomly chosen; second, 50% of items from the chosen section were randomly selected.

Parameters of Divergence Algorithm

The number of random subsets in Step 1 (parameter m) was equal to 100. The significance levels were $\alpha_1 = 0.001$, $\alpha_2 = 0.1$, and $\alpha_3 = 0.001$. To prevent over-detection of examinees, the Bonferroni correction (Abdi, 2007) was applied, i.e., the significance level α_3 was divided by 4 (because the LSAT has four operational sections). Also to prevent the overdetection, the Divergence Algorithm ignored test centers where less than two examinees were detected (i.e., it was assumed that an aberrant subgroup had at least two examinees). Random subsets in Ω and subsets generated in the simulated annealing were constrained to have size from five to 20 items because each operational section of the LSAT was approximately 25 items. In simulated annealing, $P(1) = P(2) = P(3) = 1/3$, $t_0 = 100$, $d = 0.97$, and $t_1 = 0.0001$.

Performance Measures for Detecting Affected Test Centers

The Type I error rate was computed as follows (this was an empirical probability for a test center to be falsely detected):

$$\varepsilon_G = \frac{[\text{number of detected test centers}] - [\text{number of correctly detected test centers}]}{[\text{number of unaffected test centers}]} \quad (8)$$

The detection rate was computed according to the following:

$$\beta_G = \frac{[\text{number of correctly detected test centers}]}{[\text{number of affected test centers}]}. \quad (9)$$

Performance Measures for Detecting Compromised Items

The Type I error rate was computed as follows:

$$\varepsilon_s = \frac{[\text{number of detected items}] - [\text{number of correctly detected items}]}{[\text{number of uncompromised items}]} . \quad (10)$$

The detection rate was computed according to the following:

$$\beta_s = \frac{[\text{number of correctly detected items}]}{[\text{number of compromised items}]} . \quad (11)$$

Performance Measures for Detecting Aberrant Examinees

The Type I error rate was computed as follows (this was an empirical probability for an examinee to be falsely detected):

$$\varepsilon_j = \frac{[\text{number of detected examinees}] - [\text{number of correctly detected examinees}]}{[\text{number of nonaberrant examinees}]} . \quad (12)$$

The detection rate was computed according to the following:

$$\beta_j = \frac{[\text{number of correctly detected examinees}]}{[\text{number of aberrant examinees}]} . \quad (13)$$

The precision was computed according to the following:

$$\rho_j = \frac{[\text{number of correctly detected examinees}]}{[\text{number of detected examinees}]} . \quad (14)$$

Experiment 1

The number of affected test centers was five, each with one aberrant subgroup consisting of 10 aberrant examinees. The resultant performance measures are provided in Table 8.1.

Experiment 2

The number of affected test centers was five, each with one aberrant subgroup containing 10 aberrant examinees or 2 aberrant subgroups each containing 5 aberrant examinees. The resultant performance measures are provided in Table 8.2.

Table 8.1 Resultant Performance Measures

	ε_G	β_G	ε_s	β_s	ε_J	β_J	ρ_J
Divergence Algorithm	0.0	1.0	0.0	0.98	0.000	0.98	1.0
lz	NA	NA	NA	NA	0.001	0.00	0.0

Table 8.2 Resultant Performance Measures

	ε_G	β_G	ε_s	β_s	ε_J	β_J	ρ_J
Divergence Algorithm	0.0	1.0	0.01	1.0	0.000	0.88	1.0
lz	NA	NA	NA	NA	0.001	0.00	0.0

Discussion of the Results

The Divergence Algorithm detected over 85% of the aberrant examinees with a Type I error rate of approximately 0.0; in contrast, statistic lz, with Type I error rate of approximately 0.001, detected approximately 0% of aberrant examinees (see Tables 8.1 and 8.2).

The Divergence Algorithm detected over 95% of the compromised items with a Type I error rate of approximately 0.01. Also, the algorithm detected 100% of the affected test centers with a Type I error rate of approximately 0.0 (see Tables 8.1 and 8.2).

REAL DATA EXPERIMENTS

The common licensure dataset was used to demonstrate the efficacy of the Divergence Algorithm. The original dataset was split in two datasets, based on test form. Dataset Form 1 included 1,636 examinees, 46 of whom had been flagged by the test entity. Dataset Form 2 included 1,644 examinees, 48 of whom had been flagged by the test entity. Both datasets included responses to 170 items in one operational section. Four separate studies were performed with each dataset where the following profile variables were used to partition examinees into disjoint groups: variable *Country* was used for the first study; variable *State* was used for the second study; variable *School* was used for the third study; and variable *Test center* was used for the forth study.

The two abovementioned approaches to detect item preknowledge were compared. Critical values for both methods were computed from simulated data based on the item parameters provided by the test sponsor and a simulated $N(0, 1)$ examinee pool.

Parameters for the Divergence Algorithm were as follows. The number of random subsets in Step 1 (parameter m) was equal to 100. The significance levels were $\alpha_1 = 0.001$, $\alpha_2 = 0.1$, and $\alpha_3 = 0.0001$. Random subsets in Ω and subsets generated in the simulated annealing were constrained to have size from $5 \times 4 = 20$ to $20 \times 4 = 80$ items (this choice was due to four times more operational sections in the above simulation study; however, a good choice of the bounds on compromised subset is a topic of future research). In simulated annealing, $\mathbf{P}(1) = \mathbf{P}(2) = \mathbf{P}(3) = 1/3$, $t_0 = 1,000$, $d = 0.99$ (more iterations of the simulated annealing than in the above simulation study), and $t_1 = 0.0001$.

For each dataset, first, the Divergence Algorithm was applied. Next, lz was computed. To ensure comparable Type I error rates across the different methodologies, the

Table 8.3 Analysis of Form 1

Partition	Divergence Algorithm			lz		
	<i>Detected</i>	<i>Detected flagged</i>	<i>Precision</i>	<i>Detected</i>	<i>Detected flagged</i>	<i>Precision</i>
Country	109	13	0.12	109	7	0.06
State	146	16	0.11	146	8	0.05
School	62	2	0.03	62	4	0.06
Test center	33	2	0.06	33	2	0.06

Table 8.4 Analysis of Form 2

Partition	Divergence Algorithm			lz		
	<i>Detected</i>	<i>Detected flagged</i>	<i>Precision</i>	<i>Detected</i>	<i>Detected flagged</i>	<i>Precision</i>
Country	26	5	0.19	26	4	0.15
State	24	2	0.08	24	4	0.17
School	16	0	0.00	16	3	0.19
Test center	17	1	0.06	17	3	0.18

flagging criteria for lz were set to identify the exact same number of aberrant examinees as was found using the Divergence Algorithm. The following characteristics were computed:

- #1 number of detected examinees (referred as *Detected*)
- #2 number of detected flagged examinees (referred as *Detected flagged*)
- #3 *Precision*, computed as characteristic #2 divided by characteristic #1.

The results are shown in Tables 8.3 and 8.4.

Discussion of the Results

The performance of the Divergence Algorithm depends on the variables used to partition a given dataset into groups (see Tables 8.3–8.4). The number of detected examinees was high with a large variance across different partitions and datasets (see Tables 8.3–8.4). Although the test sponsor is quite confident that flagged examinees were aberrant and flagged items were compromised, it is certainly possible that there were other aberrant examinees and compromised items that were not flagged by the test sponsors but were actually identified by one or both of the methods used here.

SUMMARY

In general, item preknowledge is hard to detect due to multiple unknowns involved: unknown subgroups of examinees (from unknown affected groups [e.g., affected test centers]) accessing unknown compromised subsets of items prior to taking the test (see Figure 8.1). The objective of this chapter was to identify all three unknowns by

methods of information theory and combinatorial optimization. This is an important task because if each compromised subset can be identified accurately, then it is possible to (a) detect examinees with preknowledge of those items with high power (Belov, 2015) and (b) remove those items (and only those items) from the test bank so that future examinees are not unfairly advantaged.

The major result of this chapter is formulated as the Divergence Algorithm. This algorithm is an extension of the 3D Algorithm by Belov (2014). The Divergence Algorithm demonstrated impressive performance on simulated data (see Tables 8.1 and 8.2). In particular, it was able to identify compromised subsets with high precision. On real data, the Divergence Algorithm was able to identify multiple examinees (but not all of them) flagged by the sponsor (see Tables 8.3 and 8.4).

This chapter should be interpreted as a work in progress because of the following:

1. The first step of the Divergence Algorithm (detecting affected groups) is crucial for its overall performance, because if an affected group is undetected, then all its examinees will skip further steps of detection. The Divergence Algorithm employs a random search, which might converge too slowly when the operational section is larger (such as in real dataset used here) and/or compromised item subsets are smaller. More research is needed on this step of the algorithm.
2. The Divergence Algorithm is an algorithmic framework where embedded subroutines and statistics can be modified to improve overall performance for a specific testing program. The following modifications are possible:
 - 2.1 The Divergence Algorithm can be applied for CAT, MST, and CBT with posteriors of speed (see van der Linden [2011] for details on response time modeling). In this case, the following divergence statistics can be used:

$$\mathbf{D}^*(\mathbf{F}_U \parallel \mathbf{F}_C) + \mathbf{D}^*(\mathbf{V}_U \parallel \mathbf{V}_C), \quad (15)$$

where \mathbf{V}_C is the posterior of speed computed from response times to the items from C and \mathbf{V}_U is the posterior of speed computed from response times to the items from U . The statistic (15) uses additional information from examinees response times, which should improve the overall performance of the Divergence Algorithm.

- 2.2 Instead of using simulated annealing within the Divergence Algorithm, the following alternatives for the combinatorial search of the compromised subsets could be used: a greedy heuristic (Papadimitriou & Steiglitz, 1982); a genetic algorithm (Mitchell, 1996); or a tabu search (Glover & Laguna, 1997).

REFERENCES

- Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 103–107). Thousand Oaks, CA: Sage.
- Armstrong, R. D., Stoubos, Z. G., Kung, M. T., & Shi, M. (2007). On the performance of the lz person-fit statistic. *Practical Assessment Research & Evaluation*, 12(16). Retrieved from <http://pareonline.net/pdf/v12n16.pdf>
- Belov, D. I. (2013). Detection of test collusion via Kullback-Leibler divergence. *Journal of Educational Measurement*, 50, 141–163.
- Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing*, 2(3), 37–58.
- Belov, D. I. (2015, online first). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, 40, 83–97.

- Belov, D. I., Pashley, P. J., Lewis, C., & Armstrong, R. D. (2007). Detecting aberrant responses with Kullback–Leibler distance. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 7–14). Tokyo: Universal Academy Press.
- Bertsimas, D., & Tsitsiklis, J. (1993). Simulated annealing. *Statistical Science*, 8(1), 10–15.
- Choe, E. (2014). *Utilizing response time in sequential detection of compromised items*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67–86.
- Glover, F. W., & Laguna, M. (1997). *Tabu search*. Boston, MA: Kluwer Academic Publishers.
- Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220, 671–680.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 79–86.
- Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, 53, 161–176.
- Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Mitchell, M. (1996). *An introduction to genetic algorithms*. Cambridge, MA: The MIT Press.
- Obregon, P. (2013). *A Bayesian approach to detecting compromised items*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- O’Leary, L. S., & Smith, R. W. (2013). *Extending differential person and item functioning to aid in maintenance of exposed exams*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Papadimitriou, C. H., & Steiglitz, K. (1982). *Combinatorial optimization: Algorithms and complexity*. Englewood Cliffs, NJ: Prentice-Hall.
- Shu, Z., Henson, R., & Leucht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item comprise. *Psychometrika*, 78(3), 481–497.
- van der Linden, W. J. (2011). Modeling response times with latent variables: Principles and applications. *Psychological Test and Assessment Modeling*, 53, 334–358.
- van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73(3), 365–384.

9

USING RESPONSE TIME DATA TO DETECT COMPROMISED ITEMS AND/OR PEOPLE

Keith A. Boughton, Jessalyn Smith, and Hao Ren

INTRODUCTION

Another example of how technological capabilities have enhanced opportunities for improving test integrity is the potential for computer-based testing programs to capture information on the amount of time examinees take to respond to test items. A number of studies have examined the use of these data (response latency) for detecting various forms of cheating on tests, particularly item preknowledge and item harvesting.

In this chapter, the authors summarize the research on using response time data to detect various forms of test cheating. Additionally, the chapter identifies best practices in the use of response latency data, applies those models to the common dataset(s), and introduces a refined model that addresses limitations in current approaches.

It is important to acknowledge the centrality of maintaining test integrity and security for the validity of inferences or decisions made on the basis of test results. Test security issues are a growing concern as the stakes associated with the test results increase. Maintaining the integrity of the test is critical for both financial and practical reasons. The loss of a set of items through overexposure or through collusion can represent not only a significant loss financially but also has serious implications for the psychometric integrity of the reported test scores and on the interpretations and consequences of those scores as cautioned in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014; Standard 8.7).

This chapter will use statistical procedures for identifying aberrant or irregular response behavior by the test takers, which requires response times for each item, examinee responses to each item, and item parameters. This chapter will focus on the identification of aberrances in testing that indicate possible cheating, such as preknowledge of some of the items. Specifically, we will use Bayesian procedures for identifying aberrant responses and response time patterns in testing for examinees that test online. This approach uses a combination of a regular response model and a response time (RT) model in a hierarchical framework that allows us to use the collateral information

in the responses to make the detection of aberrant response times more powerful, and the other way around (van der Linden and Guo, 2008).

Some Background on RT Research

Past research on detecting irregular behavior by test takers has largely focused on item or person misfit, at least in item response theory (Meijer and Sijtsma, 1995). As discussed in van der Linden and Guo (2008), there is much research that has been done on person-fit analyses—specifically in regard to residual procedures that focus on individual responses—that is similar to the RT procedures. The authors then go on to discuss the issue around the analyses using residuals to detect aberrant item or person responses; however, they note that using residual analyses in this way can be problematic when the item difficulty and examinee ability are similar. They then go on to note that this case of items and ability matches will occur often in adaptive testing, which means that residual analyses will lose their power to detect aberrances. Given this, van der Linden and Guo (2008) introduced the idea of using response times (RTs) as a variable to use in the detection of aberrant examinee behavior.

RT can be helpful because (1) RT is a continuous variable, which is helpful in statistical testing and the determining the size of aberrances; (2) if test items match the ability of examinees, statistical tests can still discriminate between both likely and unlikely response and thus maintain their power; and (3) RTs are a result of how quickly a test taker responds to items, compared to how much work is actually required by the items (i.e., time intensities). For this last point, it is important to note that if we have an RT model that allows us to separate out test-taker speed and time intensities, then we can adjust the test-takers' RTs for their speed and then assess whether or not they follow the pattern of time intensities. Even if an examinee tries to cheat by adjusting their RTs, it would be unlikely that they would know what the typical pattern of time intensities are (van der Linden and Guo, 2008).

Van der Linden and Guo (2008) extended the RT model introduced in van der Linden (2006). Van der Linden (2006) showed that there are two distinctly different approaches to modeling RTs, with one approach using an item response theory (IRT) model for the response variables for the same items, and the second approach modeling these RTs independently of the response variables for the items. However, the RT model in van der Linden (2006) was based on a third approach that assumes response and RT distributions are determined by distinct parameters with the statistical relationship that is captured by a second level of modeling (van der Linden, 2007). In addition, the extension is also derived from a hierarchical framework used in van der Linden (2007) for the analysis of speed and accuracy on the test items that is closer to how examinees may operate in an actual testing situation. He notes that some experimental reaction time research equates the time with the speeds at which a person operates and an assumption that times have identical distributional forms for a given person across tasks. Van der Linden (2007) then goes on to argue that these ideas of reaction time should not be used in the context of RTs on test items. Van der Linden (2007) then presents this hierarchical framework for the analysis of speed and accuracy, which consists of an IRT model, a model for response time distributions, and a higher level structure that is used to account specifically for the dependences between the item and ability parameters. Van der Linden and Guo (2008) then extends the RT model in van der Linden (2006), using a hierarchical framework from van der Linden (2007) for analyzing speed and accuracy, which uses collateral information (i.e., the

examinee response vector) on the test-takers' speed, allowing one to predict the RT on a given item, from the RTs on the other items on the test, as well as their joint RTs and responses. The extended model has both an RT model and a regular IRT model, as first and second levels for both item and person parameters. The extended model permits statistical inferences to be made about the relationship between an examinee's response time and the pattern of time intensities of the items without specifying a relationship between the observed responses and the corresponding RTs. Overall, this accounts for differences in item difficulty and time intensity in a way that may be more realistic of an examinee's testing experience.

RESPONSE TIME MODELS

This chapter applies two approaches in RT modeling: (1) the lognormal RT model as defined by van der Linden (2006) and (2) a hierarchical RT model as presented by van der Linden and Guo (2008). The second model uses a combination of an RT model with a regular response model in a hierarchical framework to detect aberrant response times, in which collateral information on the test-takers' speed is derived from their response vectors.

A general lognormal RT model models a fixed person speed, τ , for an examinee with i items. It should be noted that the larger the value of τ , the shorter amount of time is spent on an item.

$$f(t_{ij}; \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}[\alpha_i(\ln t_{ij} - (\beta_i - \tau_j))]^2\right\} \quad (1)$$

where t_{ij} is the response time of person j on item i . β_i is the time intensity parameter for item i , with higher values corresponding to an item being more time intensive for the examinee. In this model, α_i can be interpreted as a discrimination parameter with larger values corresponding to less variance in the log response times on item i for the test takers.

The hierarchical lognormal RT model (van der Linden and Guo, 2008) is specified using two levels. Level 1 defines separate models for the item responses and the RTs, with Level 2 defining the population parameters for Level 1 of the model. The item response model can be specified as any IRT model. The response-time model in Level 1 is the same as lognormal model defined in Equation 1. Level 2 of the model defines a bivariate normal distribution for the examinee ability (θ) and response speed (τ):

$$(\theta, \tau) \sim mvn(\mu, \Sigma), \mu = (\mu_\theta, \mu_\tau), \Sigma = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}. \quad (2)$$

It should be noted that both the lognormal and hierarchical models can be fit using the Rasch IRT model.

METHODS

The two response time models were fit using Bayesian sampling (i.e., Markov chain Monte Carlo; MCMC) estimation with a Gibbs sampler using the software programs WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) and R. The prior distributions

of α , β , a , and b were set using the recommendations from van der Linden (2006; van der Linden and Guo, 2008).

$$\begin{aligned} \alpha &\sim \gamma(2, 4/\text{var}(\log(t))) \\ \beta &\sim N(\text{mean}(\log(t)), 1/(\alpha^2)) \\ a &\sim \text{logN}(0, 2) \\ b &\sim N(0, 0.5) \\ (\tau, \theta) &\sim \text{MVN}((0, 0), \Sigma), \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{aligned} \quad (3)$$

The posterior distributions of α , β , a , and b were approximated for all parameters. Parameter estimates were obtained based on 3,000 draws after 5,000 preconvergence iterations.

METHODS TO FLAG ITEMS AND PERSONS

It is important to note that certain types of test behavior may show aberrances for many reasons, and thus it is useful to review both the model fit as well as RT patterns. For example, there might be memorization of items during a test or sharing of information across examinees after a test, resulting in preknowledge of a subset of the items on a subsequent test. Patterns of RTs that are not similar to the time intensities may result from things like memorization, with incorrect responses, or preknowledge may be revealed as a combination of unlikely RTs and correct responses.

We examined the posterior predictive probabilities to look at unusual RT patterns for both of these situations. Following the methods used by van der Linden and Guo (2008), the posterior distribution of τ_j given the observed RT and response pattern was obtained to identify unusual RT patterns. The lower tail predictive p -value was calculated using the posterior defined as:

$$p = \int_{-\infty}^{t_{ij/i}} f(\tilde{t}_{ij/i} | t_{ij/i}) d\tilde{t}_{ij/i}, \quad (4)$$

where $t_{ij/i}$ is the observed RT and $f(\tilde{t}_{ij/i} | t_{ij/i})$ is the posterior predicted density. The observed log RTs were standardized using the mean and variance of the predictive distribution. A band at the conventional values of +1.96 and -1.96 was used identify unexpected RT patterns. Another approach to summarizing the RT flagging is by using graphical displays, as demonstrated by van der Linden and Guo (2008). The graph can be used to represent one examinee's response pattern for each item that an examinee was given, with the residual log RT plotted next to the observed RT. Note that this type of graphical display may help explain the testing behavior of a flagged examinee.

DATA SOURCE

The data examined in this chapter is the common dataset provided from a licensure exam with two forms. Each form contains 170 operational items as well as 10 items in a pretest set. At this point, we have only looked at the 170 operational items for each form. In addition, it is important to note that examinees with a RT of zero were removed from the analyses. There were 12 examinees removed from Form 1 and 15

removed from Form 2. Given this, we used a total of 1,624 and 1,629 examinees for Forms 1 and 2, respectively.

RESULTS

Lognormal RT Model Results

Results implementing the lognormal response time data indicated that the model fit well. To verify model fit, the cumulative distribution of the left-tail posterior predictive probabilities of the observed log RTs over examinee. As is shown in Figure 9.1, most of the items left-tail posterior predictive probabilities fell along the identify line, which is a good indication of overall model fit for the lognormal RT. The item parameters used to fit the model were those given in the common licensure dataset.

Flagging Persons Using the RT Model

To identify aberrant response patterns, each log RT was standardized using the predicted mean and standard deviation given the RTs for all of the items taken by the same test taker. However, it doesn't seem realistic to think that a testing program would investigate an examinee with only a single RT flagged. Because the method for flagging is based on item-level data, it is most likely that a number of examinees will have at least one item flagged. Therefore, for the purpose of this chapter, we chose to flag individuals if an examinee was at or above the 95th percentile in terms of number of items for which the examinee's RT were flagged. For both forms, an examinee at the 95th percentile had 13 or more items with RTs flagged. Figure 9.2 displays the number of items with flags by the number of examinees for Forms 1 and 2.

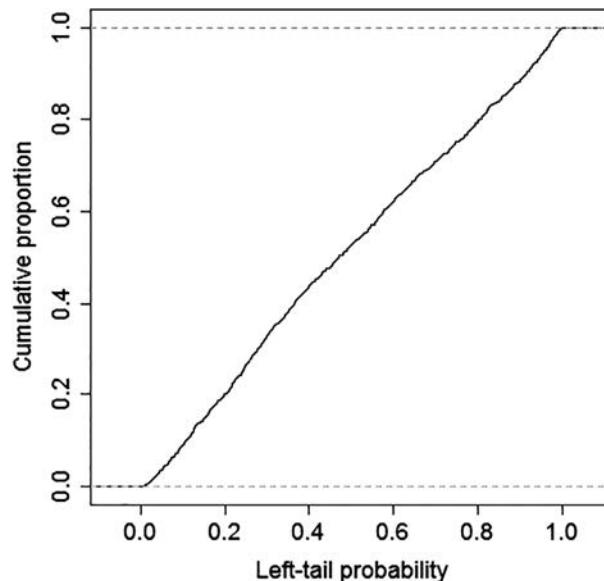


Figure 9.1 Item Example of the Cumulative Distribution of the Left-Tail Posterior Predictive Probabilities of the Observed Log RTs

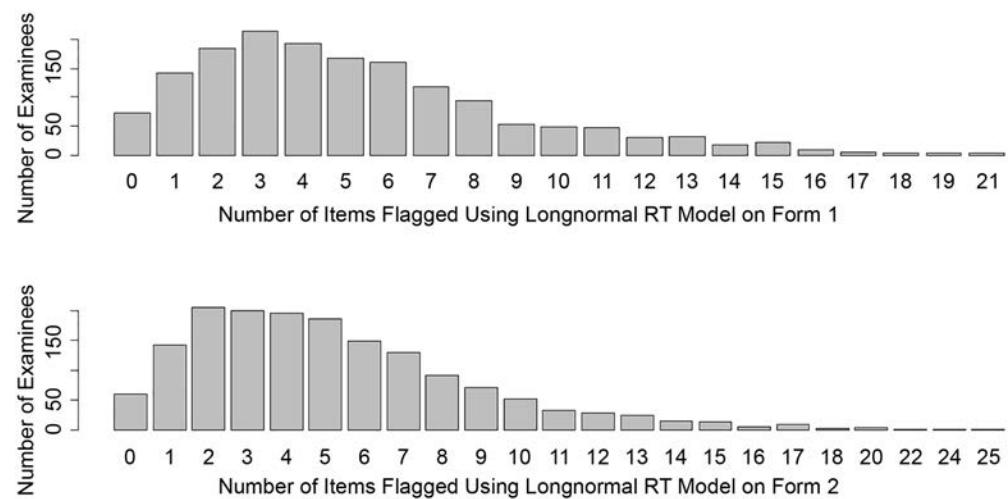


Figure 9.2 The Number of Items Flagged by the Number of Examinees for Forms 1 and 2 Using the Lognormal RT Model

Table 9.1 Examinees Flagged on at Least 13 Items Using the Lognormal RT Model and/or the Testing Company

		Flagged in the Common Dataset	
	Flagged RT Model	No	Yes
Form 1	No	1508	37
	Yes	79	5
Form 2	No	1500	34
	Yes	83	7

Of the 1,624 examinees that took Form 1, 90 were flagged as having aberrant RTs that took less time than expected, for at least 13 items. Of the 1,629 examinees that took Form 2, 84 were flagged as having aberrant RTs that took less time than expected, for at least 13 items. Note that seven of the examinees from Form 1 were flagged by both the RT model and in the common licensure dataset. Similarly for Form 2, five individuals were flagged by both the RT model and the testing company. Table 9.1 displays a summary of flagged individuals.

After flagging examinees, one can use a graphical display of the residual RT and the observed response time for every item examinees takes to see if they may have been flagged due to a testing affects (e.g., running out of time) that could explain the flagged RT results and may not be the result of a testing impropriety. Figure 9.3 is an example plot from an examinee from Form 2 who was flagged for 25 flagged RTs and who also took the exam four or more times. The x -axis represents the item in the order the examinee took them, with the y -axis representing the standardized residual RT. The white columns in the figure show the actual RT in minutes. As can be seen, items 95 and 113 were most likely known to the examinee (i.e., preknowledge). In addition, we added Figure 9.4 showing RT patterns for one examinee with more than 17 items flagged with aberrant RT patterns. Figure 9.5 shows a portion of the aberrant RT pattern; however, on further inspection, it looks as though the examinee may have run out

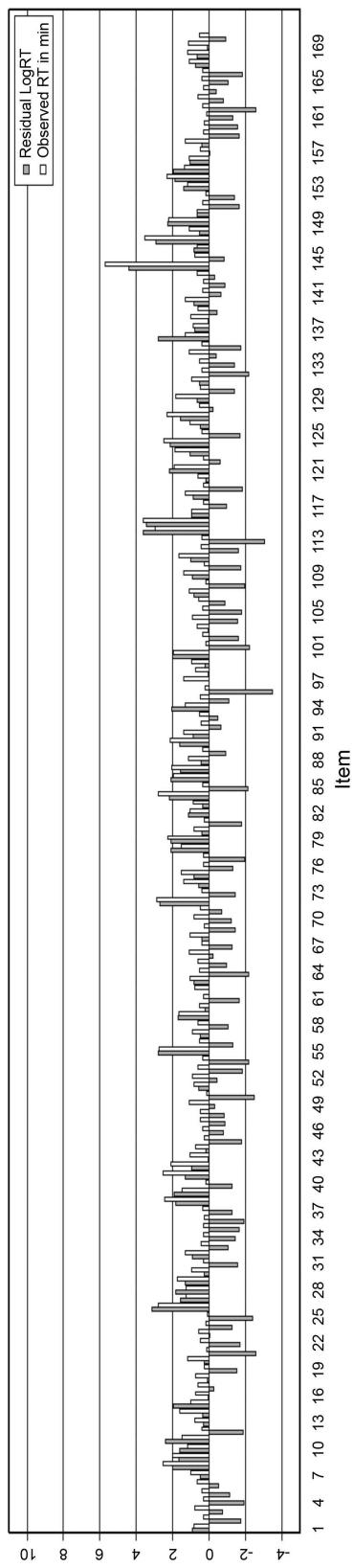


Figure 9.3 Aberrant RT Pattern of a Flagged Examinee (residuals are shaded)

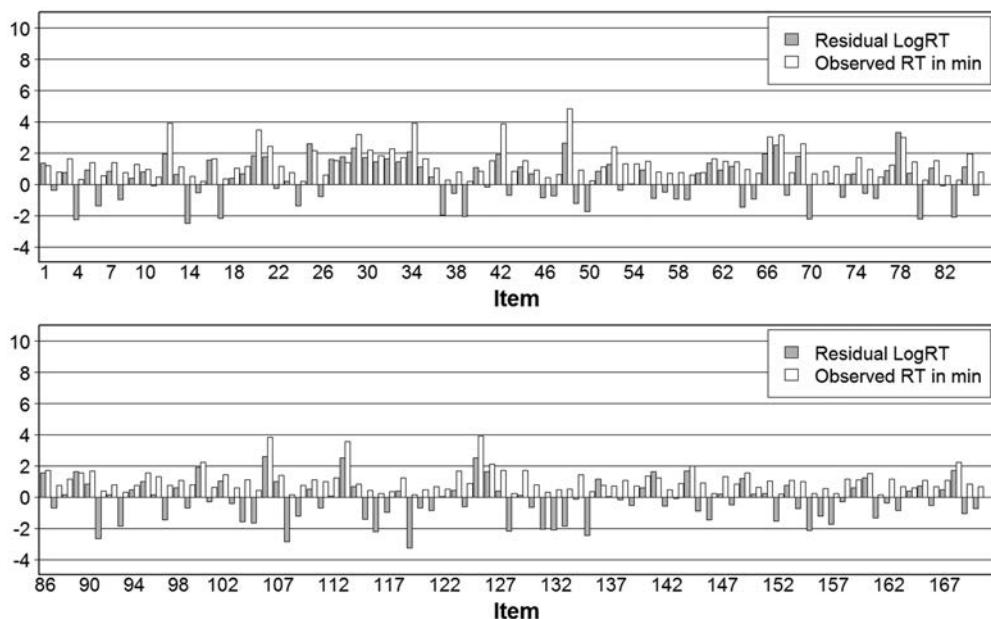


Figure 9.4 Residual Log RT Plots for Form 2 Part 1: Examples of Aberrant RT Behavior for Examinees with More Than 17 Items Flagged (residuals are shaded)

of time (i.e., speededness) as opposed to any cheating, given that the larger and negative RTs for the last several items on the test. Figure 9.6 shows a regular RT pattern for a test taker taking Form 1.

Flagging Items Using the RT Model

We also examined the number of items that had multiple examinees' RTs flagged (while conditioning on ability). Table 9.2 displays the number of items associated with a percentage of examinee RTs flagged. For example, there were two items on Form 1 that had more than 5% of the examinee RTs flagged. A cutoff would need to be specified to determine which items would need further investigation. Treating item flagging similar to how we flagged examinees, items were flagged for review if they had the number of examinee RTs at or greater than the 90th percentile of the distribution. The 90th percentile was selected because we would in practice select more items to ensure they were not exposed. For Forms 1 and 2, that would correspond to having 68 and 69 examinee RTs flagged. Of the 64 items flagged on Form 2 by the testing company, eight were flagged by our criteria. Similarly, of the 61 items that were flagged on Form 2 by the testing company, 11 were flagged using our criteria.

Table 9.3 summarizes the flagging results for items. Results indicate that 13 of the flagged items on Form 1 were also on Form 2, and nine of the flagged items on Form 2 appeared on Form 1. Nine of the flagged items appeared on both Forms 1 and 2. This would indicate that 10% of the common items were flagged.

Last, we examine the flagged examinees and the percentage of RTs flagged for each item to determine if items should be flagged for review. For both forms, there were no clear patterns of RT flagging by item for the examinees flagged in the earlier section. The percentage of examinee RTs flagged from the flagged examinee group averaged around 1.4%.

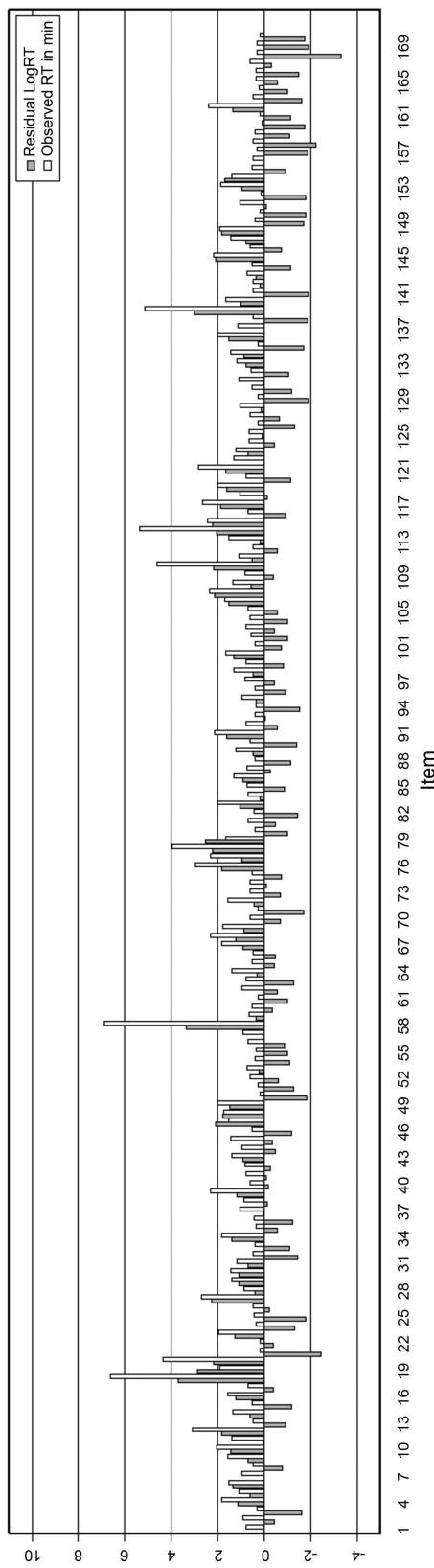


Figure 9.5 Residual Log RT Plots for Form 1: Example of Speededness (residuals are shaded)

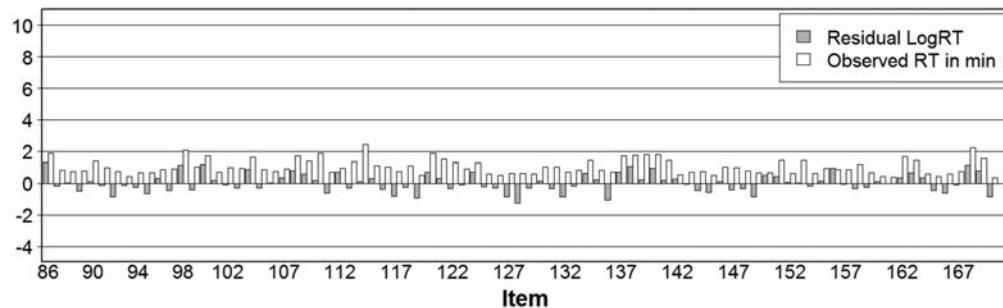


Figure 9.6 Residual Log RT Plots for Form 1 Part 2: Example of a Regular RT Pattern

Table 9.2 Items Flagged Using the Lognormal Model

% Examinees with Flagged RT	No. of Items	
	Form 1	Form 2
0–1	5	3
1–2	21	25
2–3	49	46
3–4	73	64
4–5	20	31
>5	2	1

Table 9.3 Items That Were Flagged Using Lognormal RT Model and/or in the Common Licensure Dataset

	Flagged RT Model	Flagged in the Common Dataset	
		No	Yes
Form 1	No	95	56
	Yes	11	8
Form 2	No	98	50
	Yes	11	11

Hierarchical RT Model Results

For the hierarchical RT model, results similar to those reported for the lognormal RT model are presented below. Model fit for the hierarchical RT model was acceptable using cumulative distribution of the left-tail posterior predictive probabilities of the observed log RTs. As in Figure 9.7, most of the items left-tail posterior predictive probabilities fell along the identify line, which is a good indication of overall model fit for the hierarchical RT.

Table 9.4 gives a statistical summary of the item parameter estimates' posterior means and standard errors associated with those estimates for the hierarchical model time intensity (β) and discrimination (α) parameters.

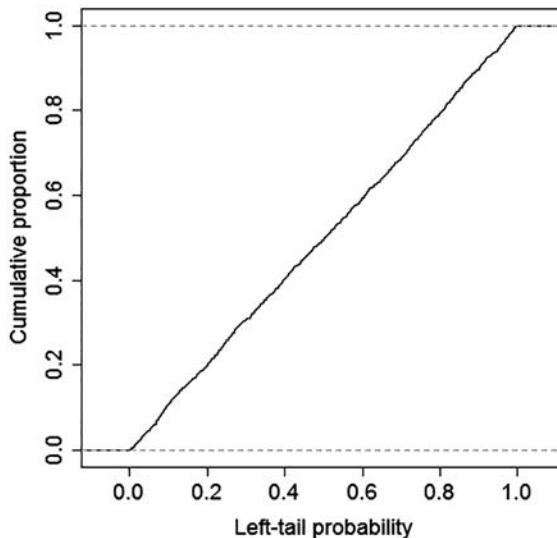


Figure 9.7 Item Example of the Cumulative Distribution of the Left-Tail Posterior Predictive Probabilities of the Observed Rts Using the Hierarchical Model

Table 9.4 Distribution of Posterior Means and *SDs* of the Item Parameters in the Hierarchical RT Model

		α		β	
		Mean	SD	Mean	SD
Form 1	Mean	2.04	0.00	3.94	0.04
	Min	1.37	0.02	2.78	0.05
	Max	2.73	0.03	4.84	0.05
	<i>SD</i>	0.28	0.00	0.33	0.00
Form 2		α		β	
		Mean	SD	Mean	SD
	Mean	2.09	0.00	3.97	0.06
	Min	1.41	0.02	2.83	0.06
	Max	2.74	0.04	4.84	0.06
	<i>SD</i>	0.29	0.00	0.33	0.00

Additionally, we note that the correlation between latent ability, θ , and response time, τ , was estimated to be .33 for Form 1 and .35 for Form 2. This indicates that higher ability examinees tend to respond slightly more quickly.

Flagging Persons

Using the same reasoning as the for the flagging criterion for the RT model, for the hierarchical model we chose to flag individuals if an examinee was at or above the 95th percentile in terms of number of items an examinee has flagged under the hierarchical model. For both forms, an examinee at the 95th percentile had 16 items or more items

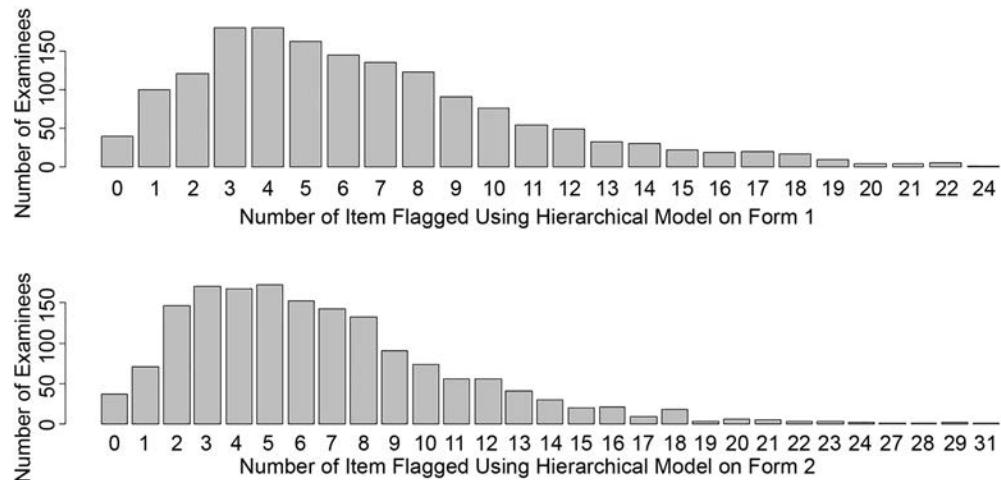


Figure 9.8 The Number of Items Flagged by the Number of Examinees for Forms 1 and 2 Using the Hierarchical Model

Table 9.5 Examinees Flagged on at Least 15 Items Using the Hierarchical Model and/or the Testing Company

	Flagged Hierarchical	Flagged in the Common Dataset	
		No	Yes
Form 1	No	1,507	35
	Yes	76	6
Form 2	No	1,498	36
	Yes	89	6

flagged on Form 1 and 15 or more items flagged on Form 2. Figure 9.8 displays the number of items with flags by the number of examinees for Forms 1 and 2.

Of the 1,624 examinees that took Form 1, 82 were flagged as having aberrant RTs that took less time than expected, for at least 16 items. Of the 1,629 examinees that took Form 2, 89 were flagged as having aberrant RTs that took less time than expected, for at least 15 items. Note that for each form, six of the examinees were flagged by both the RT model and in the common licensure dataset. Table 9.5 displays a summary of flagged individuals.

Flagging Items Using the Hierarchical Model

Table 9.6 displays the number of items associated with a percentage of examinee RTs flagged using the hierarchical model.

Using the same criterion for flagging items as was used with the lognormal RT model, items were flagged for review if they had the number of examinee RTs at or greater than the 90th percentile of the distribution under the hierarchical model. For Forms 1 and 2, that would correspond to having 70 examinee RTs flagged. Of the 64 items flagged on Form 1 by the testing company, six were flagged by our criteria, and of the 61 items that were flagged on Form 2 by the testing company, eight were flagged using our criteria. Table 9.7 summarizes the flagging results for the hierarchical model.

Table 9.6 Items Flagged Using the Hierarchical Model

% Examinees with Flagged RTs Using the Hierarchical Model	No. of Items	
	Form 1	Form 2
0–1	5	3
1–2	19	24
2–3	48	40
3–4	73	65
4–5	23	37
>5	2	1

Table 9.7 Items That Were Flagged Using Hierarchical Model and/or in the Common Dataset

	Flagged by RT Model	Flagged in the Common Dataset	
		No	Yes
Form 1	No	96	58
	Yes	10	6
Form 2	No	100	53
	Yes	9	8

Of the 16 items on Form 1 that were flagged using the hierarchical model, 11 were also on Form 2. Out of the 17 items on Form 2 that were flagged using the hierarchical model, seven appeared on Form 1. Seven of the flagged items appeared on both Forms 1 and Forms 2. This would indicate that about 10% of the common items were flagged.

DISCUSSION AND RECOMMENDATIONS

It is important to acknowledge the need of maintaining test security to support the validity of the inferences that will be made based on the test scores. This chapter utilized a lognormal RT, as well as a Hierarchical RT model, to identify aberrant or irregular response behavior by the test takers, which requires response times for each item. We focused on the identification of aberrances in testing that indicates possible cheating, such as preknowledge of some of the items. Examining how the models performed is a difficult task. From prior work, these models have been demonstrated to be efficient and effective models at detecting response time anomalies and item exposure when looking at response time. The model detected some of the same items and examinees flagged by the licensure company. Furthermore, we selected arbitrary cutoffs because of simplicity of the application. This cutoff was chosen because we felt that using this method would not only be defensible but also easily explained to those outside of the field.

CONCLUSIONS

When looking at these results and comparing them to what the licensure company found, we can report only a modest amount of common items and persons that were flagged when using response time modeling with the lognormal RT model or the hierarchical

model. Note that although the licensure company is confident that the flagged individuals had preknowledge, it is certainly possible that there are some undetected individuals remaining in the dataset. In addition, other explanations for students that we identified could be due to things like poor time management. Given this, it is important to note that the point of this research was to use a variety of methods to triangulate the results to get a better understanding of the dataset. Because it is entirely possible that the licensure company discovered the cause of the breach and, through the investigative process, was able to discover others involved, it is not necessarily the case that item compromise was evident through examination of these individuals' item responses and latency patterns. The models only point out unusual behavior with respect to response time, and it is clear that other kinds of behavior may have contributed to the irregularities and resulted in true instances of concern. This supports our belief that when investigating testing irregularities, multiple sources of evidence should be collected and assessed to draw final conclusions. Additionally, we feel that these sources of evidence should consist of observational reports, any reported testing issues, and multiple statistical methods—all of which should align to industry practices.

RT analysis can be a valuable tool used to look at test security; however, future research should look at the current RT analysis methods and evaluate how they perform in the context of identifying classroom/school-level test security breaches, as is done currently with erasure analyses.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, Inc.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—a Bayesian modelling framework concepts, structure, and extensibility. *Statistics and Computation*, 10, 325–337.
- Meijer, R. R., & Sijtsma K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, 8, 261–272.
- R Core Team. (2014). R: A language and environment for statistical computing (Version 3.1.2) [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.
- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2), 181–204.
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287–308.
- van der Linden, W. J., & Guo, F. M. (2008). Bayesian procedures for identifying aberrant response time patterns in adaptive testing. *Psychometrika*, 73(3), 365–384.

Section IIc

Detecting Unusual Gain Scores and Test Tampering



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

10

DETECTING ERASURES AND UNUSUAL GAIN SCORES

Understanding the Status Quo

Scott Bishop and Karla Egan

INTRODUCTION

This chapter provides a contextual background for the chapters that follow in this section. Those chapters cover the use of quantitative methods for detecting two different indicators of cheating: (1) frequent wrong-to-right (WR) answer changes, either by test takers during testing or, more commonly, by external manipulation of answer documents on paper-and-pencil tests, and (2) large score gains from one test administration to another, which is commonly accepted as evidence that inappropriate actions (e.g., cheating-related WR erasures) may have taken place. Like other section introductions, this chapter

- reviews the history of the methods used to detect these threats,
- describes the nature of these threats with respect to the integrity of test scores,
- defines unusual erasures and gains scores,
- reviews the application of several quantitative methods to detect these threats, and
- introduces important issues that will be addressed in subsequent chapters.

HISTORY AND CURRENT STATUS

Experienced practitioners in large-scale K–12 settings will likely recall the Lake Woe-begone reports, where Cannell (1987, 1988, and 1989) questioned the integrity of achievement gains being made in all states on norm-referenced tests. Cannell concluded that achievement gains often resulted from unethical teaching practices at the local level and lax test security measures at the state level. The security-related research that followed Cannell’s reports considered causal factors (e.g., how *teaching to the test* could result in score gains) instead of methodologies for detecting erasures and score gains. It took more than a decade before large-scale assessment programs began looking at erasures and gain scores.

Erasure Analyses

Researchers have found that students rarely erase their original responses. Qualls (2001) initially reported this finding. Primoli, Liassou, Bishop, and Nhouyvanisvong (2011) found that erasures occur in roughly one out of every 50 items, whereas Egan and Smith (2014) found that elementary students typically erased one item per test. There is also evidence that erasures occur with increased frequency on more difficult items and with mid- to high-ability examinees (Mroch, Lu, Huang, & Harris, 2014; Primoli et al., 2011).

Certain associations have emerged in prior research. Cheating that involves erasures and large score gains typically occurs at the group level, such as in classrooms and schools. Indeed, a relatively high percentage of erasures in an entire state can be traced to a relatively low percentage of schools (Primoli et al., 2011). Additionally, Primoli (2014) found that schools with a long history of not meeting adequate yearly progress tended to have higher rates of erasures than other schools.

Wibowo, Sotaridona, and Hendrawan (2013) looked at the feasibility of using item-level information (as opposed to examinee or examinee group information) on wrong-to-right erasures to support the removal of items from operational use. They proposed a statistical test to detect item compromise based on wrong-to-right erasures. If an item has evidence that it has been compromised, then it may warrant retiring the item from the existing item pool.

Simon (2014) used a data-mining approach to identify schools with suspicious erasure behavior. Using that approach, a peer group of schools is first created, and then schools are flagged as outliers relative to their peer group. Simon concluded that this approach is more sensitive than traditional approaches for detecting schools that are outliers with respect to their erasure frequency.

Gain Scores

Jacob and Levitt (2003, 2004) provide an early, well-known attempt for using gain scores to detect teacher cheating. They proposed a method based on a combination of two indicators: (1) unexpected test score fluctuations and (2) unusual student answer patterns. The first indicator is of primary interest here. Jacob and Levitt used a relatively simple method of ranking classroom-level gain scores and comparing the rankings across two time points. Plackner and Primoli (2014) used principal components analysis to study 10 different data forensics methods and concluded that Jacob and Levitt's score-fluctuation index was the second most influential in terms of accounting for variation among the components.

Since Jacob and Levitt's work, researchers have explored other ways of examining gain scores. Skorupski and Egan (2011, 2014) used Bayesian Hierarchical Linear Modeling to model change in individual scores, nested within groups, using both real and simulated data. With this approach, unusually large group-by-time interaction effects indicated possible aberrance (unusual score gains).

Liu, Liu, and Simon (2014) examined gain scores using a Bayesian polynomial mixed-effect growth model. By creating retrospective growth norms based on 5 years of cohort data, they were able to examine how student's year-to-year growth deviated from the expected growth norm. They compared this approach to the simpler, more common practice of comparing scale scores across 2 years. The two methods identified different schools, and neither method appeared superior. Even so, the authors

concluded that the Bayesian approach was the more sensitive method because it looks at positive and negative growth patterns at the student level.

Although gain score analysis is often based on comparing aggregate scale scores, some methods compare changes in performance levels instead. Clark, Skorupski, and Murphy (2013) looked at classroom-level gain scores using cumulative logit regression to predict test-takers' performance-level categories from their previous year's test scores within the same content area. Classrooms were flagged by comparing the classrooms' expected proportion of students in each achievement level to the observed proportion. This method had good detection power when cheating was not widespread in the simulated population.

Current Practice

Many testing programs now routinely use analysis of erasures and scores gains as part of their wider security protocols. Cannell (1989) reported that only six states used erasure analyses even though "the technology (was) widely available and inexpensive" (p. 22). (Cannell did not consider the analyses of gain scores as a method of detecting security breaches.) By 2013, 33 states reported using erasure analyses, and 28 states reported using gain score analyses to detect possible aberrant test-taking behaviors in their K–12 assessments (www.gao.gov/assets/660/654721.pdf).

DEFINING CONCEPTS

Erasures

On paper-and-pencil tests, an *erasure* occurs when a bubbled-in response on an answer document is erased and a different response is bubbled in. There are two legitimate reasons why a student's test response may change: rethinking and misalignment (Korts, Mead, & Bishop, 2011). *Rethinking* involves a student changing his or her mind about which answer option is correct. A *misalignment* occurs when a student is working on an item (say Item 10) in his or her test booklet but inadvertently marks a response in an option bubble for Item 11 on the answer document. All responses that follow on the answer document would be shifted down by one item until the mistake is noticed and corrected. Misalignments can result in multiple consecutive erasures where there is usually a discernable lag or lead pattern between erased and marked responses across items.

Erasures also occur because of testing irregularities. For example, a student who decides to copy off another student may need to erasure previously marked answers to bubble in the copied responses. However, the cause of many recent cheating scandals in K–12 large-scale testing programs has been the paper-and-pencil test format's susceptibility of having students' answer documents manipulated after the student has submitted his or her responses to artificially improve test scores. Unfortunately, without further evidence, researchers cannot say with certainty what actually causes high erasure rates.

Types of Erasures

Different types of erasures are illustrated in Table 10.1. Wrong-to-right (WR) erasures are of the greatest interest when evaluating possible cheating. Of course, right-to-wrong (RW) and wrong-to-wrong (WW) erasures also occur. In Table 10.1, Items 2–4

Table 10.1 Examples of Marks and Erasures That Can Occur on Answer Documents

Key	Item	Option				Comments
		A	B	C	D	
A	1.	●	○	○	○	No erasure with a dark response that completely covers bubble
A	2.	●	○	●	○	Wrong-to-Right Erasure (WR)
D	3.	○	○	●	●	Right-to-Wrong Erasure (RW)
C	4.	○	●	○	●	Wrong-to-Wrong Erasure (WW)
A	5.	●	○	○	○	WR with a very light erasure mark
D	6.	○	●	○	●	WR with still darker erasure
A	7.	●	●	○	○	WR with darker erasure that could be called a double grid (DG)
D	8.	○	●	○	●	True DG (same darkness gradient and coverage in each bubble)
B	9.	●	●	○	○	WR with only partial key coverage.
D	10.	○	∅	○	●	Stray Mark that could be miscoded as a WR
C	11.	○	○	○	○	Omitted or not reached item

illustrate these types of erasures. It is also possible for a student to erase a response, then bubble in the same response again. This type of *answer-changing behavior* cannot be detected on paper-and-pencil tests, but it can be tracked by many computer-testing platforms.

Optical Scanners

The processing of answer documents typically occurs with a dedicated optical mark recognition (OMR) device or with an image scanner that is supported with OMR software. As an answer document is processed, detailed information about the marks inside every answer option bubble can be recorded. Two frequent measures of interest are the darkness of the response (e.g., its value on a black-and-white gradient scale) and the area of the response bubbled that is covered (e.g., the percentage of pixels in the bubble's image that are covered). Items 5–7 show erasures that have different darkness gradients. Item 9 illustrates a response that only partially covers a response bubble.

Software programs are employed to enforce logical rules about whether a response is treated as a legitimate response, erasure, or double grid. For example, if marks are detected in multiple response bubbles for an item, the logical rules will define that the one that is the darkest and has the most complete coverage be treated as the response of record (i.e., the student's intended answer). Other marks detected by the scanner might be defined as erasures by the program's logic. For example, if there is more than one additional mark, the one with the second darkest gradient and second most complete coverage would be reported as the erased response. The logical rules are especially important in a case like Item 9, where one mark is light but completely fills its bubble, and the another mark is dark but only partially covers its bubble.

Given the information above, it may not be surprising that scanner results will often differ from human judgments about erasures. In fact, humans have been found to count more erasures than scanners (Wollack & Maynes, 2011). As an example, an erased response might be so light that a scanner will not report it as an erasure (perhaps Item 5 in Table 10.1). Alternatively, there may be a stray mark on the answer

document, such as Item 10 in Table 10.1, which could be falsely recorded as an erasure by the scanning process. In other cases, a response may be reported as a *double grid* even though it may be a poorly erased response (perhaps Items 9 and 10). Despite these challenges, scanners are the most cost-effective way of evaluating erasures (Wollack & Maynes, 2011).

Although, scanner technology is not the focus of this chapter, the information provided above indicates that some understanding of the details about how answer documents are scanned is important because different logic rules can be employed to determine how different combinations of the darkness gradients and the coverage inside response bubbles are reported. Further, when reading research studies, one should keep in mind that the analyzed data across studies may have employed different logical rules for determining erasures.

Gain Score Analyses

Gain score analyses examine the difference scores between two testing periods, with the primary intention of uncovering individuals, classes, schools, or districts with unexpectedly or improbably large score gains. As a difference score where outcomes include gains and losses, it is probably more appropriately referred to as score *fluctuation analyses*. But because systematic cheating is done to increase test performance, the interest is primarily in identifying score gains. At its simplest, gain scores can be computed when summative test results from Year 1 are subtracted from those of Year 2. For example, school-level mean scale scores in Year 1 are subtracted from Year 2, and schools with unreasonably large scale score gains are flagged.

It is important to note that a score gain analysis that utilizes scale scores will often yield a different impression of growth than a score gain analysis that uses the difference in the percentage of students at or above a proficient/satisfactory performance level (Holland, 2002). Figures 10.1–10.3 illustrate how such differences occur. In these figures, Year 1's cumulative distribution function (CDF) is the solid line and Year 2's CDF is the dashed line. Horizontal and vertical growth at three cut scores (140, 150, and 165) are illustrated. The dotted projections from the CDFs to the *x*-axis illustrate growth in scale score units at the cut scores. The lines projecting to the *y*-axis illustrate growth in terms of the percentage of students below at the cut scores.

In Figure 10.1 there is a uniform shift in the scale scores (i.e., the Year 2 scale scores are three points higher than Year 1 scale scores). At each cut there is consistent growth of three units from Year 1 to Year 2 along the score scale. In contrast, there are three different perspectives of growth along the cumulative distribution functions. The greatest growth was at a cut of 150 (about 11%), followed by 140 (about 7%) followed by 165 (about 5%). The difference in the percentage of students at or above an accountability cut score (PAC—a common index of growth in school accountability) is the *complement* of the cumulative distribution function—that is, the total distribution (100%) minus the percentage of the distribution at and below a specific score. PAC results may be dependent on the cut score values (Ho, 2008).

Figure 10.2 provides another example. Here, the Year 2 distribution is mostly shifted to the right of the Year 1 distribution (there is some slight crossing of the CDFs at top of score scale). Growth using the score scale difference is greatest at a cut of 140 (about 5 points), followed by 150 (about 3 points) followed by 165 (0 points—i.e., no growth). Growth using PAC differences is greatest at a cut of 150 (about 14%), followed by 140 (about 10%) followed by 165 (0 percent—i.e., no growth).

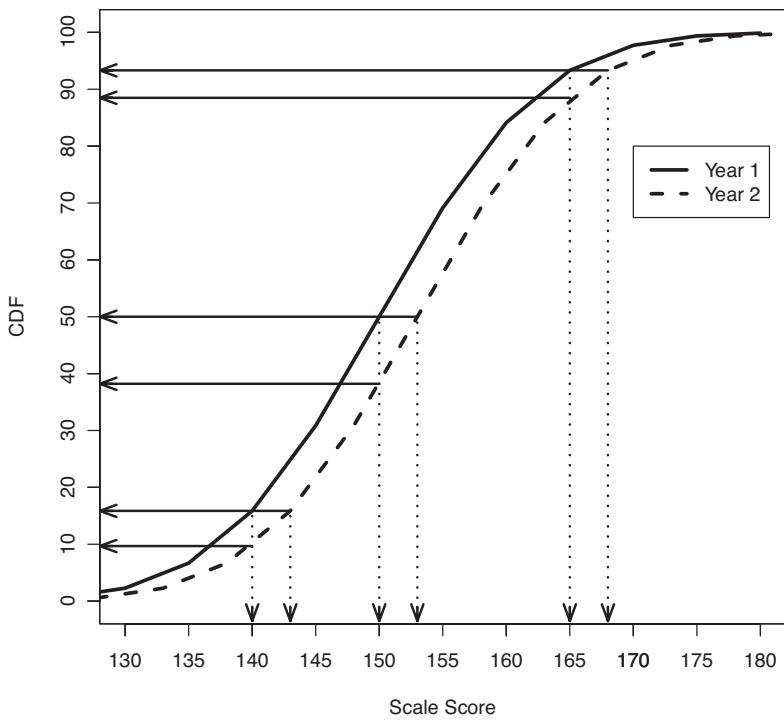


Figure 10.1 Uniform Shift in Cumulative Distribution Functions

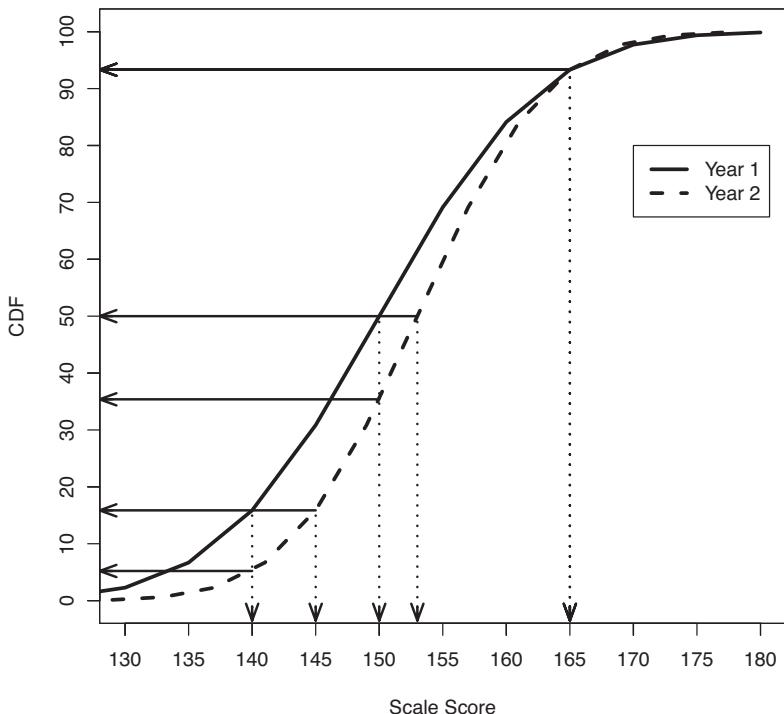


Figure 10.2 Nonuniform Shift in Cumulative Distribution Functions

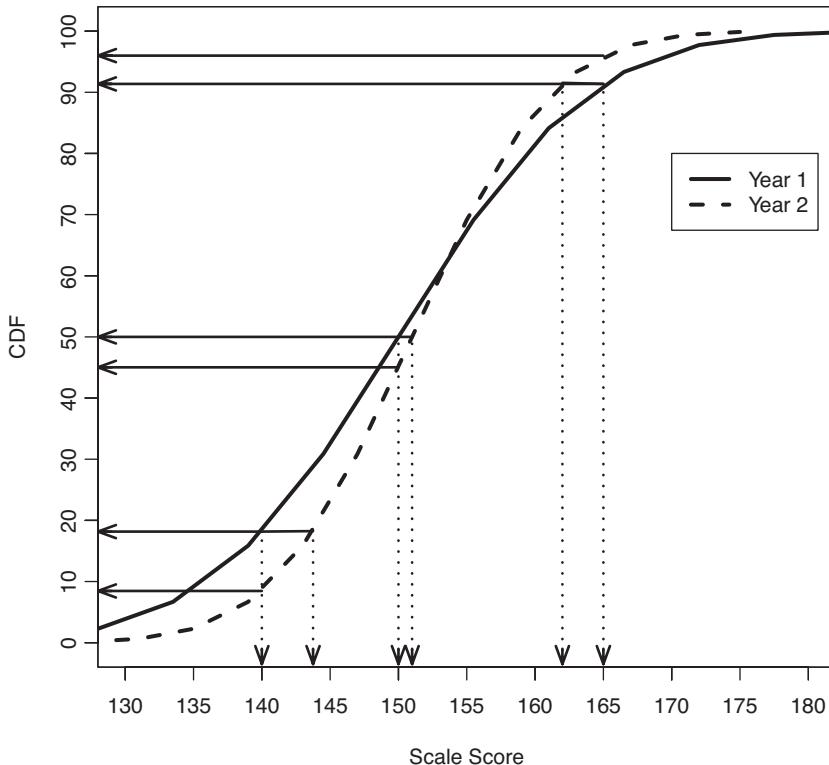


Figure 10.3 Crossing Cumulative Distribution Functions

Finally, in Figure 10.3, the CDFs cross, causing positive growth to appear for lower scale score cuts, but *negative* growth occurs for high scale score cuts. Growth using PAC differences is greatest at a cut of 140 (about 10%), followed by 140 (about 4%). At 165, the percentage of students at the cut score decreased from Year 1 to Year 2. Due to the complications of using horizontal and vertical differences in score distributions, Ho (2009) has proposed alternative measures of growth that apply nonparametric methods. However, they are not widely applied in large-scale K–12 settings at this time.

FLAGGING OBSERVATIONS

The typical end product of an erasure or gain score analysis is a list of flagged observations (classes, schools, or districts). One can brainstorm many procedures for flagging unusual observations during erasure or gain score analysis. In fact, creating a list of all the possible analyses can become somewhat overwhelming due to the following:

- The number of potential variables (dependent, independent, and covariates) involved.
- The number of ways these variables can be expressed (e.g., WR sum, WR/TE ratio).
- Application (or lack) of smoothing (Poisson, Negative Binomial, Kernel, etc.).
- The specific flagging methods employed (e.g., outliers, such as in box and whisker plots; residuals, such as from a regression analysis, significance levels from statistical tests, and ad hoc rules for various thresholds like effect-size standard deviation unit or the percentage of observed distribution selected).

The following section explores some possible methods.

DATA FRAME CONSIDERATIONS

Data files similar to the one illustrated in Table 10.2 are often available to those conducting erasure analyses. Information about the data hierarchy is needed (unique district, school, class, and student identifiers). Item-level erasure information can be included. Erasure information for 12 items is shown Table 10.2. Coding schemes are used to identify erasure types. In this example, no erasure (NE) = 0, a WR erasure = 1, a RW erasure = 2, and a WW erasure = 3. The sum over all erasure types provides the total erasures (TEs), which can range from 0 to k , where k is the number of items. Possible values for any erasure type sum can also range from 0 to k but will be restricted by the sum of other erasure types.

In some cases, more complete data from answer document processing can be made available. Some examples include the specific darkness gradient values for marks, coverage areas inside response bubbles, and pre- and posterasure responses. Additional information such as this is valuable in both applied and research settings because it can greatly expand the analyses that can be conducted. Consider several answer documents from the same class where most nonerased responses had less than 100% bubble

Table 10.2 Sample Data File Capturing Erasure Information—12 Items, 21 Students

District	School	Class	Student	Erasure String	Erasure Sums			Total Erasures
					WR	RW	WW	
1	1	1	01	000000000000	00	00	00	00
1	1	1	02	111111111111	12	00	00	12
1	1	1	03	222222222222	00	12	00	12
1	1	2	04	333333333333	00	00	12	12
1	1	2	05	100000000000	01	00	00	01
1	2	3	06	020000000000	00	01	00	01
1	2	3	07	003000000000	00	00	01	01
1	2	3	08	123000000000	01	01	01	03
1	2	4	09	110000000000	02	00	00	02
1	2	4	10	022000000000	00	02	00	02
2	3	5	11	003300000000	00	00	02	02
2	3	5	12	112233000000	02	02	02	06
2	3	5	13	100000000001	02	00	00	02
2	3	6	14	100000000002	01	01	00	02
2	3	6	15	100000000003	01	00	01	02
2	4	7	16	200000000001	01	01	00	02
2	4	7	17	200000000002	00	02	00	02
2	4	7	18	200000000003	00	01	01	02
2	4	8	19	000000000111	03	00	00	03
2	4	8	20	000000000123	01	01	01	03
2	4	8	21	100002200333	01	02	03	06

Note: No erasure (NE) = 0, WR erasure = 1, RW erasure = 2, WW erasure = 3.

coverage and lighter darkness gradient values in the response bubbles (presumably made by students), whereas most of the new responses for the erased items had complete bubble coverage and darker darkness gradient values (possibly made by school staff). Such a case would seem to provide additional circumstantial evidence that tampering with the answer document may have occurred.

MODEL FITTING

Dependent Variables

There are several ways that the erasure data may be analyzed. For example, the WRs may be summed and analyzed at the student level, or they may be aggregated and analyzed at the classroom, school, or district level. Instead of looking at the sum of the WR erasures, a researcher may choose to look at the ratio of the WRs to TEs at various levels of aggregation. This section considers some ways that erasures may be used as a dependent variable.

WR Sums

The WR sum is frequently used as a dependent variable in erasure studies. The WR sum is frequency (count) data because it represents the sum of the WRs over individual items. Because it is computed from item-level variables, those component parts are worth considering. Alone, an individual item's erasure status (NE, WR, RW, and WW) could be modeled with a multinomial distribution. When adding to achieve the WR sum, one cannot consider each item a Bernoulli trial where *success* (1) would be a WR and all other erasure types would be a *nonsuccess* (0). This is because each item has a different (although possibly similar) probability of a WR. It is also unclear if erasures across all items are independent events. Hence, fitting a binomial model to the WR sum is not appropriate.

Table 10.3 provides an example of WR sums for individual students modeled with normal, Poisson, and negative binomial (NB) distributions. Modeling individual WR sums with the Normal distribution is not prudent because it is unlikely to fit the WR sums very well, resulting in misleading interpretations. It is done here only for illustrative purposes. The inverse cumulative distribution (i.e., percentage of observations at and above) is provided for the observed data as well as for the three statistical distributions. The first distribution percentages less than 5%, 1%, and 0.1% are indicated with *, **, and ***, respectively, in Table 10.3. The z column is the value of the WR sum expressed in SD units relative to the state distribution.

If a researcher desired a significance level of 0.05 (proportion, one-tailed at the upper extreme) and assumed normality, WR sums greater than or equal to two would be flagged (as underlined in Table 10.3). In the actual data, over 7% of the students would be flagged using a sum of two as the critical value.

When modeling count data, the Poisson model is nearly always given consideration. However, the Poisson model may not fit all WR sum distributions well. There are two frequent fit issues that arise when modeling count data with the Poisson. The first is when the data is *zero inflated*, that is, there is a higher frequency of zeroes than the Poisson model would predict. The second concern is *overdispersion*, that is, the variance is greater than the mean and, under the Poisson model, the mean and the variance are expected to be equal.

Table 10.3 Partial Cumulative Distribution Table for WR Sum

WR	<i>z</i>	<i>f</i>	<i>f</i> %	Cum <i>f</i> %	Normal	Poisson	NB
8	8.12820	27	.1	0.21803	0.00000	0.00000	0.0423
7	7.06159	32	.1	0.31361	0.00000	0.00002	0.0981***
6	5.99499	52	.2	0.46892	0.00000	0.00030	0.2297
5	4.92838	83	.2	0.71682**	0.00004	0.00476	0.5446**
4	3.86178	184	.5	1.26639	0.00563***	0.06360***	1.3138
3	2.79517	511	1.5	2.79263*	0.25936**	0.68467**	3.2534*
2	1.72856	1498	4.5	7.26681	4.19436*	5.60090	8.4223
1	0.66196	5721	17.1	24.35411	25.39991	31.54475	23.9661
0	-0.40465	25327	75.6	100.00000	65.71320	100.00000	100.0000

Note: From Bishop, N. S., Liassou, D., Bulut, O., Dong, G. S., and Stearns, M. (2011). Values statistically significant at the 0.05 level (one tailed) are underlined. Values where the distributions are less than 5%, 1%, and 0.1% are indicated with *, **, and ***, respectively.

For the Poisson and NB distributions, a significance level of 0.05 (proportion, one-tailed at the upper extreme) would have flagged WR sums greater than or equal to three. A true Poisson distribution would have actually expected less than 1% of students to have WR sums in this range. The observed WR sum distribution is overdispersed relative to the Poisson. The NB is closer to the actual distribution, which is about 3%. As the WR sum increases, the Normal and Poisson extreme areas quickly approach zero. The NB is closer to the actual cumulative frequency for smaller WR sum values, but it also eventually falls under the observed results as the WR sum increases.

Table 10.3 is only one example. Although it might be representative of findings from other data sets, any theoretical distribution will sometimes fit the data better and sometimes fit the data worse. Fitting all reasonable distributions will provide the researcher with as much information as possible. As an alternative to flagging cases using critical values from theoretical distributions, researchers doing erasure research might also consider directly setting critical values using the observed WR sum distribution. For example, an organization could flag a percentage of students that they deem reasonable from practical or policy perspectives (e.g., perhaps resources only exist for investigating 1% of the testing units).

Although the previous material was intended to describe the properties of WR sums, it also raises a practical question: is student-level flagging a valuable endeavor, especially if the primarily concern is cheating at the group level? Procedures that flag groups are more reasonable; however, a procedure that first flags individual students, then uses that information to look for abnormally high clusters of flagged students within groups, would be a reasonable analytic approach. Examples of this approach are provided in Bishop et al. (2011).

WR/TE Ratio

The WR/TE ratio can be used as a dependent variable in erasure research, but it merits advanced scrutiny before doing so. Standardizing the WR sum using the TE sum has some advantages. First, Wollack and Maynes (2011) noted that the expected value of the WR/TE ratio (under null conditions) should remain unchanged. Additionally, more discrete outcomes are possible with the WR/TE ratio. The WR/TE ratio can

range from 0.0 to 1.0 in value. However, the number of unique WR/TE ratios is more restricted than some might expect. (Consider that $1/2$, $2/4$, $3/6$, $4/8$, etc. all equal 0.5.) The number of unique proportions can be determined by summing over Euler's totient function (<http://oeis.org/A000010>) up to the k -th value in the series (then adding one to that sum to account zero erasures). Using $k = 4$ items as an example, the number of unique proportions equals seven (0, 1, $1/2$, $1/3$, $2/3$, $1/4$ and $3/4$). For 45 items there are 1,081 possible ratios—from the $\binom{46}{2}$ (46 choose 2) combinations—but only 629 of those would result in unique proportions. More important, observing all possible proportions in any given erasure study is next to impossible because the great majority of examinees will have very few, if any, erasures. Consequently, there will not be as much disparity between the number of discrete outcomes from the WR sum and the WR/TE ratio as one might think.

When there are no erasures, the WR/TE ratio is undefined. Most computer programs will treat this as a missing observation. Given the fact that a high percentage of students are likely to have no erasures, treating these WR/TE ratios as missing (hereafter noted as WR/TE-M) might be questionable in terms of having an index that reflects the severity of erasures. Alternatively, one could consider giving students with no erasures a WR/TE ratio of zero (hereafter noted as WR/TE-0). Bishop et al. (2011) compared flagging results derived from WR sums against both the WR/TE-0 and WR/TE-M ratios. The WR/TE-0 ratio provided results that were more similar to the WR sums in terms of the total percentage of flagged classrooms. Despite flagging a similar percentage of classrooms, these two methods did not always flag the same classrooms. In contrast, the WR/TE-M procedure and WR sums disagreed considerably in terms of the total classrooms flagged—where the WR/TE-M always flagged more classrooms—and few classrooms were flagged by both procedures.

A final limitation of the WR/TE ratio is that it, in isolation, does not provide a good indication of the severity of the erasures. Even the lowest WR sum can have a high WR/TE ratio. As an extreme example, ratios of $1/1$ and $20/20$ would both have a WR/TE of 1, although only the latter case should raise eyebrows. Also, relatively small WR/TE ratios can have among the largest observed WR sums. In practice, one should consider the number of TEs when interpreting the WR/TE ratio. (This applies to WR sums as well.)

Wollack and Maynes (2011) recommended aggregating the WR and TE sums across all examinees in a group, and then using those totals before computing the WR/TE ratio (as opposed to calculating the ratio for individual students). They also defined the WR/TE ratio as 0 when no group-level erasures existed. The authors asserted that the WR/TE ratio does not become liberal in the face of optical scanners' undercounts of erasures (that is, the actual Type I error rate is not inflated relative to the nominal Type I error rate). Although the ratio might become too conservative (i.e., lower in its statistical power), Wollack and Maynes believed this would not impair the WR/TE ratios ability to detect egregious answer document tampering.

Class, School, and District WR Sum

Z-Test for Population Mean

Some common erasure analysis procedures are very straightforward. Ubiquitous to introductory statistics courses, the z-test for a population mean is often used to flag the mean WR sum for a group (classes, schools, or districts). The difference between the

Table 10.4 Example Using Z-Test for Population Mean to Flag Schools

School Name	School Mean WR	State Mean WR	State WR SD	School n	Z	Critical Value*	Flag
Obama Elementary	1.76	1.45	0.55	59	4.33	3.09	Yes
Bush Elementary	1.68	1.45	0.55	50	2.96	3.09	No
Clinton Elementary	1.67	1.45	0.55	55	2.97	3.09	No
Reagan Elementary	1.63	1.45	0.55	32	1.85	3.09	No
Carter Elementary	1.61	1.45	0.55	41	1.86	3.09	No
Ford Elementary	1.59	1.45	0.55	54	1.87	3.09	No
Nixon Elementary	1.55	1.45	0.55	92	1.74	3.09	No

Note: * The critical value of 3.09 is for a one-tailed test with $\alpha = 0.001$.

group's mean and a null population value (e.g., the student-level mean WR sum using all students in the testing program) is determined and then divided by the appropriate standard error. The standard error uses the standard deviation of the WR sums taken over all students, divided by the square-root of the group's n -count. The group is flagged if the test statistic is larger than a very conservative normal-deviate critical value:

$$Z_0 = \frac{\bar{x} - \mu_0}{\sigma_x / \sqrt{n}}.$$

There is an equivalent version of this statistical test that might be easier for many stakeholders to interpret. The test statistic can be rescaled so it is expressed on the scale of the mean WR sums. This is done by constructing an interval around the population value under the null hypothesis. In constructing these intervals, the standard error is multiplied by the same conservative normal-deviate value. To determine the critical value for the test statistic, one only needs to add this result to the null population value so as to keep the test one directional. If a group's mean WR sum exceeds the critical value, then the group is flagged. Table 10.4 provides an hypothetical example.

Before applying either variant of this procedure, one should consider the following. First, it makes explicit normality assumptions. Normality for a distribution of mean scores is usually justified by the *central limit theorem* (CLT). Because the probability of a WR erasure is very small, large n counts may be required before the normal distribution will fit the distribution of mean WR sums well. Additionally, the observations over which the means are calculated should be independent. Because students are nested within classes, independence is questionable (consider a teacher who advises students to review and change their answers as testing time permits versus a teacher who does not give that advice). Still, if employed as a flagging process, this approach might be reasonable as long as probability statements based on normality assumptions are treated very cautiously.

Deviations From the Mean

A different approach is to calculate the difference between the mean WR sum in a given group (class, school, or district) and the *grand mean* over all mean WR sums for the groups in the testing program. If the group's mean WR sum is far enough away from the grand mean, then the group is flagged. Here, far enough can be defined in

standard-deviation units, which comes from dividing the difference by the standard deviation of the groups' mean WR sums. The flagging criterion often utilizes four or more standard deviation (SD) units to offer additional protection against making a Type I error. To increase the statistical power of this test, one could take the null population value to be the grand mean of all groups *except* the group being evaluated.

A difference between the Z -test and the deviations from the mean approaches is that the first uses standard errors and is hypothesis-testing oriented whereas the second uses standard deviations and is effect-size oriented. Consideration of the *unit of analysis* is another difference between the two approaches. In the first case, student-level SDs are used to determine the standard error. In the latter case, the SD that is used is calculated over the group means, not individual values.

There may be temptation to take a hypothesis-testing orientation with the *deviations from the mean* approach. Here one would simply refer the calculated result,

$$Z_0 = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}},$$

to the normal distribution table.

The resulting test statistic is unlikely to be normally distributed, because each group (class, school, or district) will have different n -counts and standard deviations for their respective WR sums. Research by Bishop et al. (2011) has shown that the distributions of class, school, and district WR means are positively skewed, sometimes markedly for classes. Like the application of the Z -test, deviations from the mean can be considered a reasonable ad hoc flagging approach as long as probability statements derived from the normality assumptions (if applied) have caveats attached to them regarding their potential inaccuracy.

Joint Distributions

The two approaches described above used a single dependent variable. This section explores modeling and flagging criteria based on the conditional relationship between WRs and TEs (e.g., flagging observations that are greater than a critical value established for studentized residuals above a regression line).

Student Level

Bishop et al. (2011) found that simple linear regression can account for a substantial proportion of variance in student-level WR sums using TE sums as a predictor. However, there was considerable heteroscedasticity (nonconstant residual variance) with the conditional WR variance increasing as the TE sum increased. The observed heteroscedasticity makes one question the reasonableness of using the normal distribution to model residuals from the WR on TE linear regression.

There are constraints that the TE sum puts on the WR sum. As noted earlier, if there are no erasures, there can be no WRs. Similarly, the WR sum cannot exceed the TE sum, which has a maximum value equal to the number of test items. In a scatterplot of WR sums on TE sums, no data points would lie above the *identity line*. As the TE sum increases, the WR sum has more opportunity to vary because there are more possible values that it can take on. These conditions might facilitate the aforementioned heteroscedasticity.

Although there is a strong linear association between WR sums and TE sums, the pattern of residuals is problematic. A Box-Cox transformation may be considered to

deal with the observed heteroscedasticity, but maintaining the sum/count scale might be more intuitive for many stakeholders. An alternative model, like Poisson regression, can be considered. Bishop et al. (2011) found that Poisson regression models fit the WR-on-TE data well. When the percentages of students with linear regression and Poisson residuals greater than 1.96 were flagged, the following differences were observed. If these residuals were normally distributed, one would expect about 2.5% of the students to have residuals in this range. The linear regression flagged more than 2.5% of students. This could be expected because of the extreme heteroscedasticity present. The Poisson residuals flagged approximately 2.5% of students. Few students were jointly flagged by both regression procedures. Also, both regression modeling approaches tended to identify unique cases as compared to the univariate models using WR sums.

Group Level

Simple linear regression can be performed where the predictor is the mean of the class TE sums and the response variable is the mean of the class WR sums. Bishop et al. (2011) found that the proportion of variance accounted for using this approach was generally greater than the proportion of variance accounted for using student-level regression modeling. However, the improvement in fit was achieved at the cost of losing information about within-class variability.

Both of these regression-based approaches flagged fewer groups than procedures like the Z-test. Further, the percentage of groups jointly flagged by both procedures was relatively small. An advantage of regression approaches is that one can visually detect potentially influential data points and outliers using scatterplots. In an erasure analysis, these outliers (and perhaps points with high influence and/or leverage) could represent classrooms with aberrant WR erasures that are worth further attention.

OTHER ANALYSES

Hierarchical Linear Modeling (HLM)

Application of single-level models to multilevel data can be problematic. In addition to aggregation bias and misestimated precision (Raudenbush & Bryk, 2002), there is also missed opportunity to explore relationships that might vary across nesting groups like schools and districts. Hierarchical models avoid these problems, lead to improved parameter estimation, and provide ability to model cross-level interactions and between-groups variation. Partitioning of variance components with nested data is also possible.

In the section above regarding use of regression to model group mean WR sums, it was noted that group-level regression modeling achieved better fit than student-level modeling. However, that improved fit was achieved at the cost of losing information about within-class variability. Hierarchical linear modeling is ideally suited to analyzing erasure data due to the nesting that is inherent to the data. Bishop, Bulut, and Seo (2011) applied three candidate HLM models to WR counts. The fit of each model was obtained and contrasted against each other to see if more complex models improved fit over simpler models. The simplest model was a *two-level random intercept model within school*. Here, the mean WR counts were fit for schools. Next, a more complex model, a *two-level random slope model*, regressed the number of WRs on the individual-level TE count. The authors found that the addition of the slope parameter improved fit over intercept only model.

Next, a *three-level random slope model* was fit using students (Level 1) nested within classes (Level 2) nested within school (Level 3). The question here was if modeling variability among classes would improve fit over the slope model. The authors found that the proportion of variance of students within class (Level 1) was 0.75, whereas the proportion of variance among classrooms within schools (Level 2) was 0.08, and the proportion of variance among schools (Level 3) was 0.16. The authors' ruled out the last model because there was little evidence of class-to-class variation. The authors went on to use the two-level random slope to flag schools. Specifically, 95% confidence intervals were constructed for each school's slope, and schools were flagged if their confidence interval did not capture the overall mean slope.

There are many different ways one might apply hierarchical linear modeling (HLM) to erasure data. In the example above, a very basic linear model with normal error and identity link was used. Other link and error structures could have been considered and other predictor and criterion variables employed. (See Skorupski, Fitzpatrick, and Egan, this volume, for additional information about HLM.)

Generalized Estimation Equations

Generalized Estimation Equations (GEEs) offer a very flexible approach to modeling erasure data. Using GEEs, one can specify the well-known least squares solutions (i.e., assuming a normal error variance and an identity link function). Alternatively, simple, *semi-parametric models* can be fit where less rigid assumptions about the distribution of WRs are made.

There are two noteworthy advantages of applying GEEs. One is that the correlational structure of the data can be specified. Some options here include unstructured, independent, or exchangeable correlations. The expected covariance between nested objects (like students in classrooms) to other study variables is considered when positing a correlational structure in GEEs. If the relationship is expected to be the same for all students in classrooms, then exchangeable correlations should be posited. Or, the students within classrooms may be expected to have independent relationships with other study variables. When the relationship is uncertain, or expected to differ, unstructured correlations should be posited. Unstructured correlations may be the best choice in the erasure-analysis context but requires more intensive computation resources, especially with large data files.

Another advantage of GEEs is the ability to obtain robust standard errors (often referred to as *sandwich* or *empirically adjusted* standard errors). When models are incorrect, which to some degree they almost always are in applied practice, classical standard errors might be downward biased. If only slightly larger than their classical counterparts, the robust standard errors would be preferable under such circumstances. Further, King and Roberts (2015) note that larger differences between robust and classical standard errors can be important diagnostic indicators of serious model misspecification for researchers.

A simple model could treat district and school factors as fixed effects if one only wanted to identify classrooms with aberrantly high erasures. This would be satisfactory if there was no interest in obtaining separate parameter estimates for districts and school effects. The same would be true if only school-level results were of interest. Of course, more complex situations can be modeled. Inclusion of school and district effects would only require that they, too, have unique identifiers. One could posit designs that included nested factors, use of covariates (like TEs), and different link

(e.g., Poisson, Negative Binomial) and error (Log) functions. Fit and interpretability of various models could be compared and results from the preferred model used to flag suspicious observations.

Although there are many benefits of GEE for erasure-analysis research, there are potential limitations as well, particularly with computation and hardware/software issues. One computation issue is that the number of examinees per class would need to be sufficient to support the analysis. One might go as low as 10 students per class, but more could be preferred by some. Low ns might result in convergence problems. Depending on one's particular software, parameter estimates may not be printed, or if they are, warning messages will be attached with the results. On a more practical note, for any classrooms with smaller ns omitted from the GEE, some kind of alternative erasure analysis might still be required.

Another issue is that with more complex designs, run times can become extremely long. Having access to a computer system that has extraordinary computational capacity (i.e., a *supercomputer*) is ideal. Simplification of the design may be the only recourse (e.g., selecting an independent or exchangeable correlation structures; specifying normal variance and identity link function versus Poisson and log functions). Considerable professional judgment is recommended because inaccurate parameter and standard error estimates are likely if inappropriate models are used.

Visualization

Supporting visuals for erasure analyses are often appreciated by stakeholders. Two simple figures are provided here. Figure 10.4 shows the WR erasure proportions in 30 test

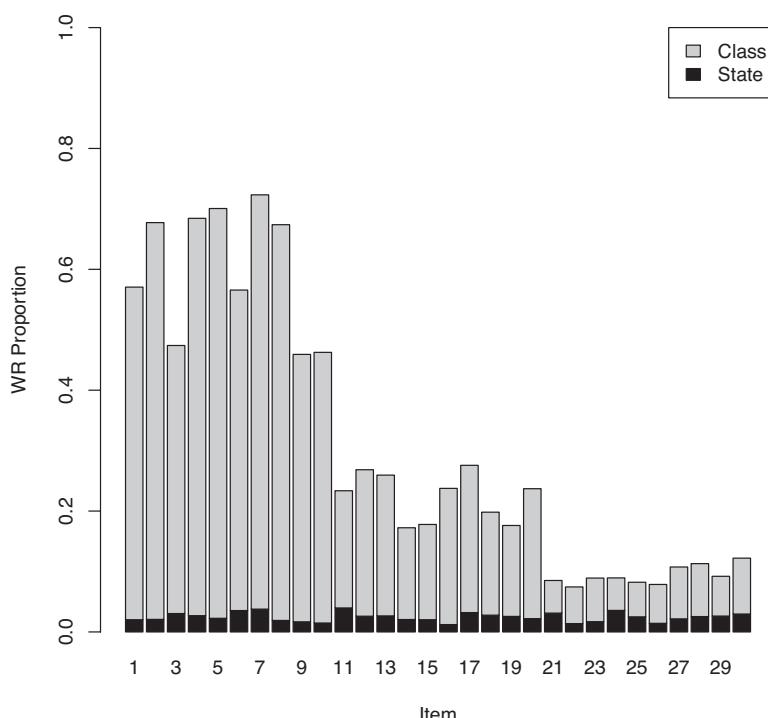


Figure 10.4 Example of State-Level Versus Class-Level WRs for 30 Items

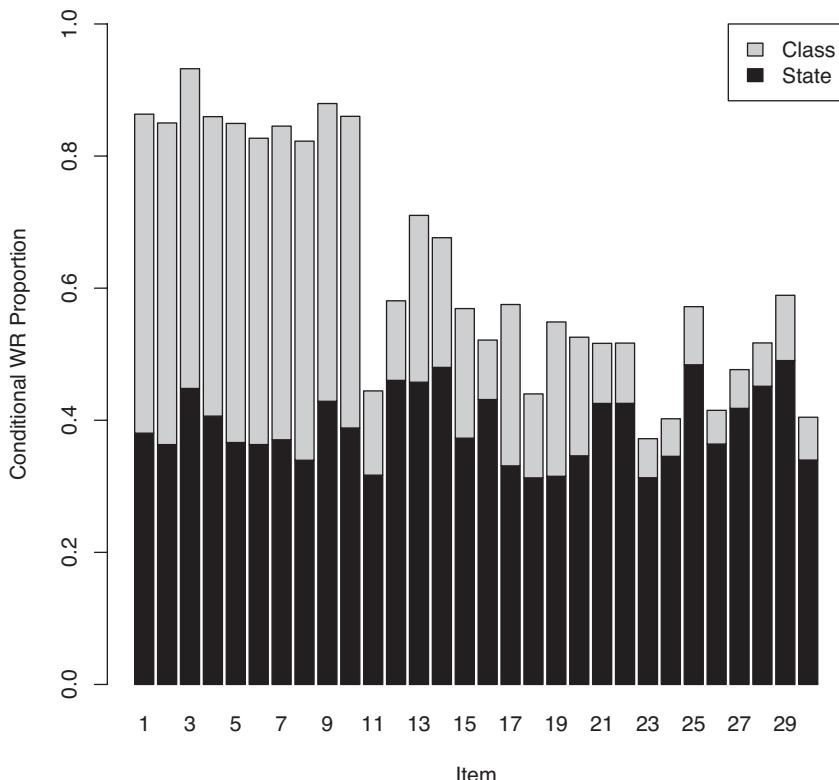


Figure 10.5 Example of State-Level Versus Class-Level Conditional WRs for 30 Items

items for students in a hypothetical classroom under investigation and the corresponding WR proportions for students in the state. For Item 1, the state had a WR proportion of just under 0.02. However, in the class the WR proportion was nearly 0.60. In this example, WRs appear to occur with much greater frequency on items at the front of the test. Perhaps additional investigation would uncover why that was so.

Figure 10.5 shows conditional WR proportions. That is, given that the item had an erasure, the probability that the erasure was WR. These conditional proportions are larger than those in Figure 10.4 and can be a nice supplement to the basic WR proportions. Specifically, stakeholders may find it telling if many items had nearly 100% of erasures resulting in WRs. Such charts can become even more informative if the items are grouped by content area, difficulty, item type, or item position, such as in the examples.

SOME RESULTS

Several of the analysis procedures described above were applied to a common data set that was made available to the chapter authors of this volume. The erasure data used below came from the Year 2, Grade 5 mathematics exam, which consisted of 54 items.

Z-Test for Population Means

Using a one-tailed alternative that the group's (class, school, or district) WR mean is greater than the overall state WR mean with an alpha of 0.001, 76 (of 3,909) classes, 35

(of 1,187) schools, and 23 (of 630) districts were flagged. As a practical matter, there is overlap across the flagged cases because a smaller district may only have one school, and a smaller school may only have one class. Here, the 35 flagged schools were nested in 30 unique districts, 21 of which intersected with the 23 flagged districts. The 76 flagged classes were nested in 67 unique schools, 32 of which intersected with the 35 flagged schools. Determining such overlap can help facilitate investigation and intervention activities.

Deviations From Mean (Effect Size)

If a *deviations from mean* flagging criteria of 0.5 standard-deviation units was used (considered by many to represent a *moderate effect size*), 350 classes, 29 schools, and 13 districts would have been flagged. Compared to the *Z-test*, more classes were flagged with the *deviations from mean* approach, but fewer schools and districts.

Joint Flagging Criteria

In a state with a large population of students, the number of classes, schools, and districts flagged by either the *Z-test* or the *deviations from mean* approach may be too large to allow the state department of education to investigate all cases thoroughly. Use of a joint flagging strategy that includes both statistical and practical significance could be beneficial in this circumstance. If a joint criterion was applied using flagged cases from both the *Z-test* and *deviations from mean*, then 71 classes, 19 schools, 7 districts would be flagged. Note that the joint flagged counts show that the *z-test* and *deviations from mean* approach do not always flag the same groups. Although there was only a modest reduction in the number of classes flagged, the flagging criteria could be modified to address this.

Group Regression

Linear regression was applied to aggregate group WR and TE means. Cases with large, positive residuals (greater than 3.09, consistent with an alpha of 0.001) were flagged. These flagged groups would have more WRs than their TEs would have predicted. The regression residual approach flagged 40 classes, seven schools, and three districts. This approach flagged fewer groups than the *z-test* results. Only 15 of the classes flagged via regression overlapped with classes flagged with the *z-test*. For schools and districts, the overlap was only three and one, respectively. Another regression approach would be to exam differences in each group's WR on TE regression line. However, the HLM approach described above will provide much richer information for this type of analysis.

Using Flagged Students to Identify Suspicious Groups

The following approach was not described above but represents an alternative flagging method. Erasures are a rare occurrence across individual students, so when clusters of students with large erasures appear within the same class or school, it is a suspicious outcome. Poisson regression of WR counts on TE counts was used to flag individual students that had large, positive residuals. Here, students with residuals greater than 1.28 (alpha = 0.10) were flagged. (A larger alpha was applied here because this is the initial step in this analysis. A more stringent alpha will be used below.) The 6.3% of

students were flagged in the state due to large residuals. If only random factors prevailed, these flagged students should be randomly spread among the classes, schools, and districts throughout the state.

Let $X = 1$ indicate that a student was flagged by the Poisson regression and $X = 0$ otherwise. Each student's individual status is essentially an independent Bernoulli trial with parameter $p = 0.063$. The sum over the Bernoulli trials, X , within a class, school, or district will have a binomial distribution with parameters n (the sample size for the group) and p . Each group's sum can be compared to the appropriate binomial distribution to determine how unlikely the group's sum is. If the outcome is significant at the 0.001 level, then the group is flagged.

Using this method, 15 classes, 11 schools, and 15 districts were flagged. Regarding overlap with the z -test results, only one class, four schools, and nine districts were flagged by both procedures. Poisson regression is not the only way one could flag individual students for the first step of this approach. One might simply flag 10% of the students with the very largest WR sums in the state. In this case, the p parameter for the binomial test would equal 0.10. Like all the analyses above, different threshold criteria would yield different results.

CONCLUSIONS/RECOMMENDATIONS

The purpose of this chapter was to review foundational material relevant to other chapters in this section. Several simple and direct analysis procedures were reviewed. Hypothesis testing procedures are often used to flag individuals or groups with high numbers of erasures. In a basic hypothesis test, a group's mean WR sum is considered a random sample from the entire state distribution of WR sum (asymptotic normality is assumed using the central limit theorem). The hypothesis tests informs stakeholder whether the group's mean WR sum is too high to be explained by random sampling. When the null hypothesis is rejected, then the group is flagged.

There are certain advantages to keeping things simple for stakeholders, who may only have a very basic understanding of statistical analysis. Practitioners must also understand that applied procedures may be covered by media and could have negative consequences for classroom teachers as well as school and district administrators. Because of the potential consequences associated with the results of these analyses, researchers will want to consider issues such as sample size, statistical power, management of Type I error rates, and suitability of methods to the available data when selecting analyses procedures. When beginning analysis for a new program, use of multiple analysis procedures can be informative. (For general advice when setting up a new analysis program, the reader is referred to National Council on Measurement in Education [2012] and Olson and Fremer [2013]).

Although there may be some benefit to conducting analyses which are accessible to the media and public at large, these are by no means the only options available to practitioners. As forensic methods become more familiar to would-be cheaters, they will look for new ways to cheat or be less blatant in the application of their typical methods. The upshot is that improved analysis procedures with more statistical power will be needed. Improved data collection (e.g., always delineating proctors/teachers versus lumping all students under a general classification, like "Grade 3") will likely be important as new analysis procedures come to light. Regardless of the accountability system (value added models, etc.), cheating will diminish the integrity of the results. The following chapters considers new methods for practitioners to consider.

REFERENCES

- Bishop, S., Bulut, O., & Seo, D. (2011, April). *Modeling erasure behavior*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Bishop, N. S., Liassou, D., Bulut, O., Seo, D. G., & Stearns, M. (2011, April). *Application and comparison of alternative procedures in erasure analysis*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average*. Daniels, WV: Friends for Education.
- Cannell, J. J. (1988). Nationally normed elementary achievement level testing in America's public schools: How all fifty states are above the national average. *Educational Measurement: Issues and Practices*, 7(2), 5–9.
- Cannell, J. J. (1989). *The "Lake Wobegon" report: How public educators cheat on standardized achievement tests*. Albuquerque, NM: Friends for Education.
- Clark, J. M. III, Skorupski, W. P., & Murphy, S. T. (2013, October). *Using nonlinear regression to identify unusual performance level classification rates*. Paper presented at the Conference on Statistical Detection of Potential Test Fraud, Madison, WI.
- Egan, K., & Smith, J. (2014). Test security for multistage tests: A quality control perspective. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 220–229). New York, NY: Routledge.
- Ho, A. D. (2008). The problem with “proficiency”: Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, 37(6), 351–360.
- Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics*, 34, 201–228.
- Holland, P. W. (2002). Two measures of change in the gaps between the CDFs of test score distributions. *Journal of Educational and Behavioral Statistics*, 27, 3–17.
- Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, 118, 843–877.
- Jacob, B. A., & Levitt, S. D. (2004). To catch a cheat. *Education Next*, 4, 69–75.
- King, G., & Roberts, M. E. (2015). How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis*, 23(2), 159–179.
- Korts, J., Mead, R., & Bishop, N. S. (2011, April). *Erasure analysis: Descriptives, covariates, and modeling options from multiple states—Coordinated session*. Paper presented at annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Liu, X., Liu, F., & Simon, M., (2014, April). *Are the score gains suspicious?—A Bayesian growth analysis approach*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.
- Mroch, A. A., Lu, Y., Huang, C.-Y., & Harris, D. J. (2014). AYP consequences and erasure behavior. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 137–148). New York, NY: Routledge.
- National Council on Measurement in Education. (2012). *Testing and data integrity in the administration of statewide student assessment programs*. Madison, WI: Author.
- Olson, J., & Fremer, J. (2013). *TILSA test security guidebook: Preventing, detecting, and investigating test security irregularities*. Washington DC: Council of Chief State School Officers.
- Plackner, C., & Primoli, V. (2014). Using multiple methods to detect aberrant data. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 203–219). New York, NY: Routledge.
- Primoli, V. (2014). An exploration of answer changing behavior on a computer-based high-stakes achievement test. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 149–157). New York, NY: Routledge.
- Primoli, V., Liassou, D., Bishop, N. S., & Nhouyvanisvong, A. (2011, April). *Erasure descriptive statistics and covariates*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Qualls, A. L. (2001). Can knowledge of erasure behavior be used as an indicator of possible cheating? *Educational Measurement: Issues and Practice*, 20(1), 9–16.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Simon, M. (2014). Macro level systems of statistical evidence indicative of cheating. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 83–100). New York, NY: Routledge.
- Skorupski, W., & Egan, K. (2011). *A hierarchical linear modeling approach for detecting cheating and aberrance*. Paper presented at the Conference on Statistical Detection of Potential Test Fraud, Lawrence, KS.

- Skorupski , W., & Egan, K. (2014). Patterns of examinee erasure behavior for a large-scale assessment. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 121–136). New York, NY: Routledge.
- Sotaridona, L. S., Wibowo, A., & Hendrawan, (2014). I. Detection of non-independent test taking by similarity analysis. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 38–52). New York, NY: Routledge.
- Wibowo, A., Sotaridona, L. S., & Hendrawan, I. (2013, April). *Statistical models for flagging unusual number of wrong-to-right erasures*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Wollack, J. A., & Maynes, D. (2011). *Detection of test collusion using item response data*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

11

DETECTING TEST TAMPERING AT THE GROUP LEVEL

James A. Wollack and Carol A. Eckerly

As a result of many high-profile incidents involving educators altering answer sheets to improve the test scores for their students, districts and states are becoming more proactive about looking for evidence of test tampering. In addition, because many states and districts have not policed themselves to the satisfaction of the general public, it is now common for local and national media to conduct their own analyses to identify potential test tampering after test scores are released. However, despite the importance of detecting test tampering and the national attention that educator cheating has received, the fact remains that even within the field of data forensics, which is itself relatively new, erasure detection is an emerging research area in which there is widespread disagreement over how best to conduct such analyses. There are few established methods and little understanding of how well these approaches work.

Overwhelmingly, the most common approaches to detecting test tampering involve comparing the numbers of erasures (or certain types of erasures) within a class to the average number of erasures statewide. Many of these approaches have focused specifically on wrong-to-right (WTR) erasures, in which an answer is changed from an incorrect answer to a correct answer. As discussed by Wollack and Maynes (2014), erasures, including WTR erasures, are a perfectly normal part of paper-and-pencil testing and are not, in and of themselves, indicative of tampering. However, erasures not attributable to tampering are generally rare events (Bishop, Liassou, Bulut, Seo, & Bishop, 2011; Maynes, 2013; Mroch, Lu, Huang, & Harris, 2012; Primoli, 2012; Primoli, Liassou, Bishop, & Nhouyvanisvong, 2011; Qualls, 2001), with baseline rates having been estimated to occur with around 2% of all items. Consequently, high erasure counts are easily identified as anomalous. Furthermore, with paper-based tests, every instance of test tampering will invariably result in erasures, especially of the WTR variety.

Most erasure-based methods to detect test tampering are based on empirical distributions of counts of total or WTR erasures (Bishop et al., 2011; Maynes, 2013; Primoli, et al., 2011). These methods are discussed in greater depth by Bishop and Egan (this volume) and were used to detect test tampering in most of the highly publicized incidents of educator cheating, so they are certainly not without merit. However, building

empirical distributions of erasure counts in the presence of test tampering will lead to spurious estimates of the false positive rate. Also, reliance on empirical distributions—and contaminated ones at that—leads to an inability to accurately quantify the unusualness of atypically large test statistics.

More recently, two papers have appeared presenting model-based approaches to detecting examinees with an unusual number of WTR erasures (van der Linden & Jeon, 2012; Wollack, Cohen, & Eckerly, 2015). In addition, both models were themselves modeled after established indexes. The van der Linden and Jeon index was modeled after the generalized binomial test (*GBT*; van der Linden & Sotaridona, 2006), whereas the Wollack et al. erasure detection index (*EDI*) was modeled after the ω statistic (Wollack, 1997). Both the *GBT* and ω have repeatedly been shown to demonstrate strong control of the nominal Type I error rate and are two of the most powerful indexes available for detecting unusual similarity among examinees. Van der Linden and Jeon did not specifically explore the Type I error rate or power of their index, pointing instead to the already demonstrated performance of the *GBT* as evidence of its utility. Wollack et al. conducted a simulation study that not only showed strong Type I error control and power (especially among the lower ability students for whom tampering is far more likely) but also robustness to several typical types of benign erasures.

Although the van der Linden and Jeon index and the *EDI* are both significant steps forward in terms of the scientific rigor of erasure detection, one significant limitation to both is that they are designed as indexes to identify individual students with unusually high erasure counts. No matter how unusual a particular student's erasure pattern may be, in practice, it is very unlikely that schools, districts, or states are going to be interested in identifying individual students. Instead, an element of the test-tampering hypothesis is that educators are tampering with answer sheets for multiple students. Yet, as of now, these indexes have not been applied to detect tampering among groups of students, significantly limiting their utility in operational settings.

The purpose of this study is to extend the *EDI* to detect tampering at the class, school, or district level. In this chapter, we will provide a brief description of the *EDI*, present an extension of this index that easily allows for its application to group-level data, and conduct a simulation study to demonstrate the Type I error rate and power of the group-based *EDI* in a variety of contexts. We will close by applying the *EDI* to the common K-12 dataset, for purposes of identifying classes, schools, and districts with unusual patterns of erasure.

DESCRIPTION OF THE EDI

For each examinee j , consider partitioning the items into two nonoverlapping sets, such that E_j is the set of items erased by examinee j and \bar{E}_j is the set of items for which j did not record an erasure. The statistic of interest in computing *EDI* is X_{j,E_j} , the number correct score for examinee j among the erased items. Under the assumption that the test consists of items for which only one correct answer exists (hence, no right-to-right answer changes exist), this sum is equivalent to the WTR erasure count. *EDI* models the WTR erasure count using a normal approximation, with continuity correction¹ because the number of erasures is often quite small. Hence, *EDI* is computed as

$$\frac{X_{j,E_j} - E(X_{j,E_j}) - \frac{1}{2}}{SE(X_{j,E_j})}, \quad (1)$$

where

$E(X_{j,E_j}) = \sum E_j P_j(\theta)$ is the expected WTR erasure count for examinee j ,

$SE(X_{j,E_j}) = \sqrt{\sum_{E_j} P_j(\theta)(1 - P_j(\theta))}$ is the standard error associated with the WTR erasure count for examinee j ,

and $P_j(\theta)$ is the probability of examinee j correctly answering each item in E_j , which is found by fitting an appropriate item response model to the data. Furthermore, for purposes of computing $P_j(\theta)$, θ_j is estimated based on items from E_j . Wollack et al. (2015) showed that by considering only nonerased items in estimating examinee ability, contamination from test tampering is much reduced in trait estimates for tampered examinees, whereas trait estimates for nontampered examinees experienced negligible amounts of bias.

EXTENSION OF EDI TO DETECT GROUP-BASED TAMPERING

The proposed extension of *EDI* to the group setting is straightforward and requires knowledge of the group memberships prior to conducting the analysis. In a school-based setting, reasonable groups for which tampering might be explored include class, school, and district, all of which are typically known in advance.

From Equation (1), for each examinee, calculation of *EDI* involves three components, X_{j,E_j} , $E(X_{j,E_j})$, and $SE(X_{j,E_j})$. Computation of a group-level index, EDI_g , involves first computing each of these three components for all examinees individually, then aggregating each component across all examinees in the group. The aggregated totals are then used to compute *EDI*, as in (1). That is,

$$EDI_g = \frac{\sum_j X_{j,E_j} - \sum_j E(X_{j,E_j}) - \frac{1}{2}}{\sqrt{\sum_j [SE(X_{j,E_j})]^2}}.$$

Under this formulation, EDI_g essentially treats the entire group as though it was a single student taking a very long test and computes the index over all erasures in the class. However, by aggregating in this manner, rather than by averaging *EDI* values, for example, each student is allowed to contribute differently to EDI_g . The strength of EDI_g lies in the fact that the number of items in E_j will generally be larger for tampered examinees than for nontampered examinees. Consequently, even within a class in which the majority of students were untampered, the number of erased items among tampered examinees may be larger than the number of erased items among untampered examinees. Hence, tampered students contribute much more to EDI_g than do untampered students, meaning that it should only take a relatively small number of tampered individuals per group to offset the watering-down effect caused by including untampered students in the composite measure. In addition, this aggregation strategy allows for the group-level index to yield an accurate probabilistic statement about the unusualness of any flagged groups

SIMULATION STUDY

Simulating Data and Erasures

In studying the properties of any index intended to detect test tampering through erasure detection, it is important to not only simulate fraudulent erasures but also benign

erasures, in which the erasure was caused by the student changing the answer of his or her own accord, without prompting. Wollack et al. (2015) described three types of benign erasures. Random erasures refer to situations in which students initially mark one answer, then reconsider their choice and change responses. Misalignment erasures are those caused by students accidentally marking their answer in the wrong field, as might be the case when they record the answer for Item 2 in the field allocated for Item 3. String-end erasures are produced when students randomly fill in the items at the end of the test as a safeguard against running out of time, then change answers as time permits them to attempt the questions.

Misalignment is likely to cause unusually high erasure counts because students often continue to mismark their answer sheet for many consecutive items before finally realizing the mistake, at which point they erase all items and bubble in the intended answers. Misalignment errors are particularly important to study because they are also quite likely to produce high counts of WTR erasures, especially for higher ability students. With any item for which the student knew the answer, the student's initial response (prior to erasure) was unlikely to be scored as correct (with probability equal to one divided by the number of alternatives); however, after erasing, the item was very likely to be scored as correct, hence producing a WTR erasure. Furthermore, any item for which the initial answer happened to be correct, only those items that were changed to incorrect will be detected by the scanner as an erasure. Right-to-right erasures, in which the student erases the correct answer but then decides to keep that answer, are indistinguishable from initial marks to scanning equipment. Consequently, the number and proportion of WTR erasures figures to be quite high when a student commits a misalignment error.

String-end erasures are also capable of producing high erasure counts; however, (a) this strategy is not likely to be used by many students, (b) those who do use it likely will not use it for many items because it is wasteful of time and time is in short supply, and (c) it is likely to produce a similar pattern of erasures as would the misalignment; hence, detection indexes should perform similarly in the two circumstances. For this reason, only misalignment and random erasures were simulated here.

Data with erasures were simulated following a three-step process, as described below:

1. Complete, untampered data were simulated for a 50-item, five-alternative test under the nominal response model (NRM; Bock, 1972), using item parameters from the college-level test of English language usage used by Wollack et al. (2015). Within each condition, a total of 1,000 schools were simulated. Schools were generated to be of different quality by sampling the school ability (θ_S) from a normal distribution with a mean of 0 and a standard deviation of 0.5 (i.e., $\theta_S \sim N(0, 0.5)$). Within school, data were simulated for a single cohort of students (e.g., fourth graders). Schools were simulated to have different numbers of classes, representing small schools (one class), medium schools (three classes) or large schools (six classes). Classes were simulated to be of three different sizes, representing small classes (15 students), medium classes (25 students), or large classes (35 students). All classes within school were assumed to be drawn from the same population of students; therefore, students in all classes were simulated from a normal distribution with mean equal to the school-level mean (θ_S) and standard deviation equal to 1 (i.e., $\theta \sim N(\theta_S, 1)$).
2. Benign erasures were simulated. Every examinee was given an opportunity to produce an erasure; however, because benign erasures are uncommon, the simulation

was conducted in a way that ensured that benign erasures were rare events. Within each school, 2% of the students were randomly selected to produce misalignment erasures, whereas the remaining 98% were candidates to produce random erasures. The process for simulating these erasures is described in detail in Wollack et al. (2015). For students sampled to produce misalignment erasures, the number of misaligned items was sampled from a binomial distribution with $N = 50$ and $p = .25$, thereby producing an average of 12.5 misaligned items per student. However, because these were five-alternative items, approximately 20% of the misaligned items would result in the same answer choice being selected, hence not being identifiable as an erasure by the scanner. This simulation process resulted in these students producing an average of 10 erased items each. For each of these students, the starting point of the misalignment was determined by randomly selecting an item between Item 1 and $50 - k + 1$, where k is the number of misaligned items. The initial answer was determined by shifting the final answers one spot. If the initial and final answers were different, it was recorded as an erasure. For the remaining students who did not misalign answers, the number of randomly erased items for each student was sampled from a binomial distribution with $N = 50$ and $p = .02$. This approach produced an average of one erased item per person, but resulted in no erasures for approximately one-third of the students. The specific items which were erased were selected at random. The initial response for these items was simulated based on the conditional NRM, in which the final response (i.e., the one simulated in Step 1) is assigned a probability of zero and probabilities of selecting each remaining alternative are found based on the logit values for those choices (see Wollack et al. 2015 for more details).

3. Fraudulent erasures were simulated. Within each school, four different proportions of tampered classes were simulated: 0%, 33%, 67%, and 100%. Fully crossing these with the three levels of classes within school (i.e., 1, 3, and 6 classes) produced 12 conditions, as shown in Table 11.1. From Table 11.1, one can see that two of those conditions (33% and 67% of schools with only one class) were null; hence, these cells included no simulations. Within each of the remaining 10 cells, we simulated four different levels of the number of erasure victims (1, 3, 5, or 10), and three different levels of the number of tampered items per erasure victim (3, 5, or 10). The specific classes in which tampering occurred and the specific items on which tampering occurred were determined randomly, subject to the constraint that all tampered items would result in WTR erasures. In the event that a student's raw score was sufficiently high that it was not possible to produce the number of WTR erasures required, the student was simply given a perfect score.

Table 11.1 Classes Within Schools \times Proportion of Tampered Classes

		% of Tampered Classes			
		0%	33%	67%	100%
Classes within Schools	1	Type I error condition	Null	Null	Power condition (1 class)
	3	Type I error condition	Power condition (1 class)	Power condition (2 classes)	Power condition (3 classes)
	6	Type I error condition	Power condition (2 classes)	Power condition (4 classes)	Power condition (6 classes)

METHOD

All simulated datasets were run through MULTILOG (Thissen, 2003) for purposes of estimating θ_j from nonerased items only. Because the simulation design was not set up to mirror reality in terms of the incidence of test tampering (i.e., when tampering was simulated, all schools were simulated to include tampering), estimating item parameters from the simulated data would have introduced an unrealistic amount of contamination into the estimates. Therefore, estimation of examinees' latent traits was based on item parameters from the NRM, though in practice, a simpler, dichotomous item response model could certainly be used. Finally, EDI_g values were computed at the class and school level by aggregating across all individuals in the respective groups.

From Table 11.1, one can see that all three cells in which no classes were tampered allow for inspection of the empirical Type I error rate of EDI_g . Similarly, the remaining seven cells of Table 11.1 describe different power conditions. For ease of presentation, discussion of the seven power conditions is with respect to the number of tampered classes. Therefore, we collapsed results from the two cells producing one tampered class, as well as those from the two cells producing two tampered classes, to produce five levels of the number of tampered classes (1, 2, 3, 4, and 6). Type I error rates and power of EDI_g were evaluated at seven different α levels between .00001 and .05.

RESULTS

Results of the Type I error study are shown in Table 11.2. Because the various independent variables did not have an important effect on the Type I error rate, the data in Table 11.2 are collapsed over all three Type I error cells shown in Table 11.1. As can be seen, the EDI_g is slightly conservative in all conditions. Although Wollack et al. (2015) also found that EDI was conservative when continuity correction was used, the results here are considerably closer to nominal. This result is expected, because EDI at the individual level (as in Wollack et al.) applies continuity correction to each examinee. Over a class of 25 examinees, subtracting half an erasure from each examinee would amount to approximately 12.5 erasures which would be subtracted from the total class count. This is a huge amount given that, in null data, only one erasure is expected per student. However, in computing EDI_g , the continuity correction is applied only once, to the final class (or school) count. It is for this same reason that the school-level Type I error rates are closer to nominal than the class-level rates; with more students included in the aggregate, the correction is a much smaller proportion of the expected number of erasures. Whereas Wollack et al. found that the power of EDI was considerably lower after continuity correction was applied, because the continuity correction is very small when applied to groups, its impact on the power should be minimal.

The power of EDI_g to detect classes is shown in Tables 11.3 and 11.4. Table 11.3 demonstrates the influence of the numbers of erasure victims and tampered items for

Table 11.2 Empirical Type I Error Rates for EDI_g

Level	False positive rate (α)						
	.00001	.0001	.0005	.001	.005	.01	.05
Class	.00000	.0000	.0002	.0004	.0022	.005	.029
School	.00000	.0001	.0003	.0006	.0035	.007	.037

Table 11.3 Power to Detect Classes, Based on Numbers of Erasure Victims Per Class and Tampered Items Per Student (Class Size = 25)

# Erasure Victims	# Tampered Items	False positive rate (α)						
		.00001	.0001	.0005	.001	.005	.01	.05
1	3	0.00	0.00	0.00	0.00	0.02	0.04	0.13
	5	0.00	0.00	0.01	0.02	0.05	0.08	0.24
	10	0.02	0.05	0.10	0.13	0.24	0.31	0.53
3	3	0.01	0.03	0.07	0.10	0.22	0.30	0.56
	5	0.10	0.20	0.32	0.39	0.57	0.66	0.85
	10	0.55	0.69	0.78	0.82	0.90	0.93	0.98
5	3	0.10	0.21	0.34	0.41	0.61	0.69	0.88
	5	0.49	0.66	0.78	0.83	0.92	0.95	0.99
	10	0.91	0.95	0.98	0.98	0.99	1.00	1.00
10	3	0.73	0.86	0.93	0.95	0.98	0.99	1.00
	5	0.98	0.99	1.00	1.00	1.00	1.00	1.00
	10	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 11.4 Power to Detect Classes, Based on Numbers of Erasure Victims Per Class and Class Size (5 erased items)

# Erasure Victims	Class Size	False positive rate (α)						
		.00001	.0001	.0005	.001	.005	.01	.05
1	15	.00	.01	.02	.03	.08	.13	.32
	25	.00	.00	.01	.02	.05	.08	.24
	35	.00	.00	.01	.01	.04	.07	.21
3	15	.19	.33	.47	.54	.71	.79	.92
	25	.10	.20	.32	.39	.57	.66	.85
	35	.05	.13	.23	.29	.47	.56	.78
5	15	.64	.79	.88	.91	.97	.98	1.00
	25	.49	.66	.78	.83	.92	.95	.99
	35	.37	.54	.68	.74	.86	.91	.97
10	15	.99	1.00	1.00	1.00	1.00	1.00	1.00
	25	.98	.99	1.00	1.00	1.00	1.00	1.00
	35	.95	.98	.99	1.00	1.00	1.00	1.00

a fixed class size of 25. Table 11.4 demonstrates the influence of the number of erasure victims and class size for a fixed number of tampered items equal to 5. Visual representations of both of these tables are provided in Figures 11.1 and 11.2 for $\alpha = .001$.

Unsurprisingly, the number of erasure victims and the number of tampered items (see Table 11.3) both have strong effects on the power of EDI_g , with power increasing as both increase. The power to detect classes is low when the class contains only one erasure victim, moderate provided it contains three erasure victims, and high when answers for as many as five students are altered. Except in situations for which there are 10 erasure victims in a class, altering as few as three items results in low-to-moderate power. Altering five items produces strong power provided at least five erasure victims

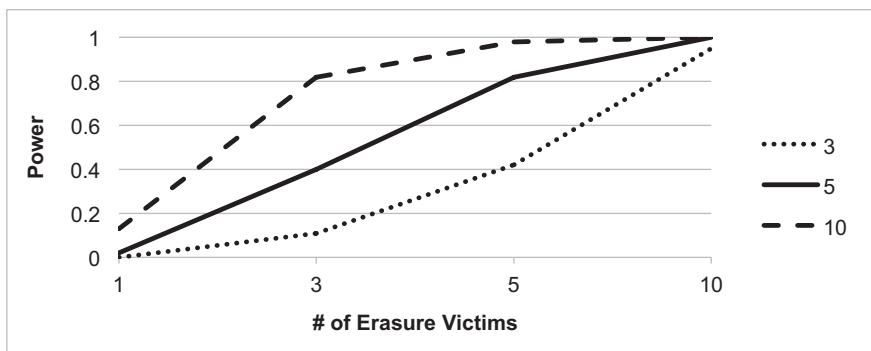


Figure 11.1 Power (at $\alpha = .001$) to detect Classes, Based on Numbers of Erasure Victims and Erased Items

Note: Class size = 25.

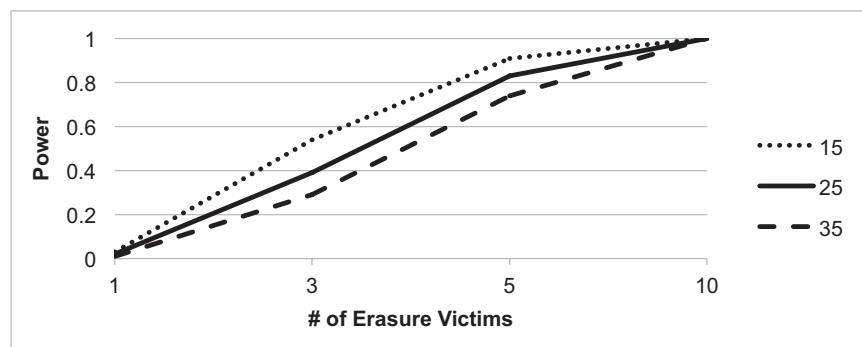


Figure 11.2 Power (at $\alpha = .001$) to Detect Classes, Based on Numbers of Erasure Victims and Class Size

Note: Number of erasures = 5.

are present. When 10 items are altered, power is low for one erasure victim, but is high for three or more. With 10 tampered items and five erasure victims, power is .90, even for α levels as small as .00001.

Class size (see Table 11.4) also has an important impact on the power, though the effect of class size is somewhat smaller than for number of erasure victims or number of erasures. Furthermore, class size is inversely related to the power of EDI_g , because for fixed numbers of erasure victims, increasing the number of students serves only to add null data to the class-level aggregate, thereby diluting the tampering effect. Although Table 11.4 suggests that the probability of detection is higher in smaller classes, this presupposes that the number of erasure victims is fixed. In practice, when educators engage in tampering, it would seem quite likely that they would alter more answer sheets in larger classes than they would in smaller classes.

The power of EDI_g to detect schools is shown in Tables 11.5 and 11.6. Table 11.5 reports on the impact of the numbers of tampered classes, erasure victims, and tampered items for a fixed class size of 25. Table 11.6 provides a similar table showing power to detect tampered schools as a function of the numbers of tampered classes, erasure victims, and class size for students with five tampered items each. Visual representations of both of these tables are provided in Figures 11.3 and 11.4 for $\alpha = .001$.

Table 11.5 Power to Detect Schools, Based on Numbers of Tampered Classes, Erasure Victims, and Tampered Items (Class Size = 25)

# Tampered Classes	# Erasure Victims	# Tampered Items	False positive rate (α)						
			.00001	.0001	.0005	.001	.005	.01	.05
1	1	3	.00	.00	.00	.00	.01	.02	.11
		5	.00	.00	.01	.01	.03	.06	.17
		10	.01	.02	.03	.05	.11	.16	.34
	3	3	.00	.01	.03	.04	.10	.16	.37
		5	.03	.07	.14	.18	.32	.40	.65
		10	.25	.39	.51	.57	.72	.78	.90
	5	3	.03	.08	.14	.18	.33	.42	.67
		5	.20	.36	.47	.52	.67	.76	.91
		10	.67	.80	.88	.90	.96	.97	.99
	10	3	.37	.55	.69	.75	.87	.92	.97
		5	.83	.91	.96	.97	.99	.99	1.00
		10	.99	1.00	1.00	1.00	1.00	1.00	1.00
2	1	3	.00	.00	.00	.01	.02	.04	.14
		5	.00	.00	.01	.02	.06	.09	.26
		10	.01	.03	.08	.11	.21	.28	.51
	3	3	.01	.03	.08	.11	.23	.32	.57
		5	.09	.20	.32	.40	.57	.65	.84
		10	.55	.70	.81	.85	.92	.95	.99
	5	3	.09	.20	.32	.39	.60	.69	.87
		5	.53	.70	.81	.86	.93	.96	.99
		10	.95	.97	.99	.99	1.00	1.00	1.00
	10	3	.78	.90	.94	.97	.99	.99	1.00
		5	.99	1.00	1.00	1.00	1.00	1.00	1.00
		10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1	3	.00	.00	.01	.02	.07	.11	.27
		5	.01	.02	.05	.08	.18	.25	.50
		10	.12	.23	.34	.41	.60	.68	.85
	3	3	.10	.22	.38	.47	.65	.73	.91
		5	.54	.73	.85	.89	.96	.98	1.00
		10	.96	.98	1.00	1.00	1.00	1.00	1.00
	5	3	.63	.81	.89	.92	.97	.99	1.00
		5	.97	.99	1.00	1.00	1.00	1.00	1.00
		10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	10	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		5	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1	3	.00	.00	.01	.02	.07	.11	.29
		5	.01	.02	.07	.09	.19	.25	.48
	3	3	.11	.21	.32	.39	.58	.67	.84
		5	.08	.20	.36	.45	.64	.73	.91

# Tampered Classes	# Erasure Victims	# Tampered Items	False positive rate (α)						
			.00001	.0001	.0005	.001	.005	.01	.05
5	1	5	.57	.75	.84	.89	.96	.97	.99
		10	.98	1.00	1.00	1.00	1.00	1.00	1.00
	3	3	.63	.79	.91	.94	.98	.99	1.00
		5	.99	1.00	1.00	1.00	1.00	1.00	1.00
	10	10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		3	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	10	5	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	1	3	.00	.01	.03	.05	.13	.21	.45
		5	.03	.07	.16	.21	.41	.52	.77
		10	.41	.59	.72	.77	.87	.91	.98
	3	3	.47	.68	.81	.86	.94	.96	1.00
		5	.95	.98	.99	1.00	1.00	1.00	1.00
	10	10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		3	.98	1.00	1.00	1.00	1.00	1.00	1.00
	5	5	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	10	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		5	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		10	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 11.6 Power to Detect Schools, Based on Numbers of Tampered Classes, Erasure Victims, and Class Size (Tampered Items = 5)

# Tampered Classes	# Erasure Victims	Class Size	False positive rate (α)						
			.00001	.0001	.0005	.001	.005	.01	.05
1	1	15	.00	.00	.01	.01	.04	.07	.20
		25	.00	.00	.01	.01	.03	.06	.17
		35	.00	.00	.00	.01	.02	.04	.15
	3	15	.07	.13	.24	.28	.45	.55	.77
		25	.03	.07	.14	.18	.32	.40	.65
		35	.01	.04	.09	.12	.24	.32	.55
	5	15	.33	.49	.63	.70	.82	.88	.96
		25	.20	.36	.47	.52	.67	.76	.91
		35	.12	.24	.37	.44	.60	.68	.85
	10	15	.93	.98	.99	.99	1.00	1.00	1.00
		25	.83	.91	.96	.97	.99	.99	1.00
		35	.73	.84	.90	.93	.97	.98	1.00
2	1	15	.00	.01	.02	.03	.10	.15	.34
		25	.00	.00	.01	.02	.06	.09	.26
		35	.00	.00	.01	.01	.04	.07	.21

(Continued)

Table 11.6 (Continued)

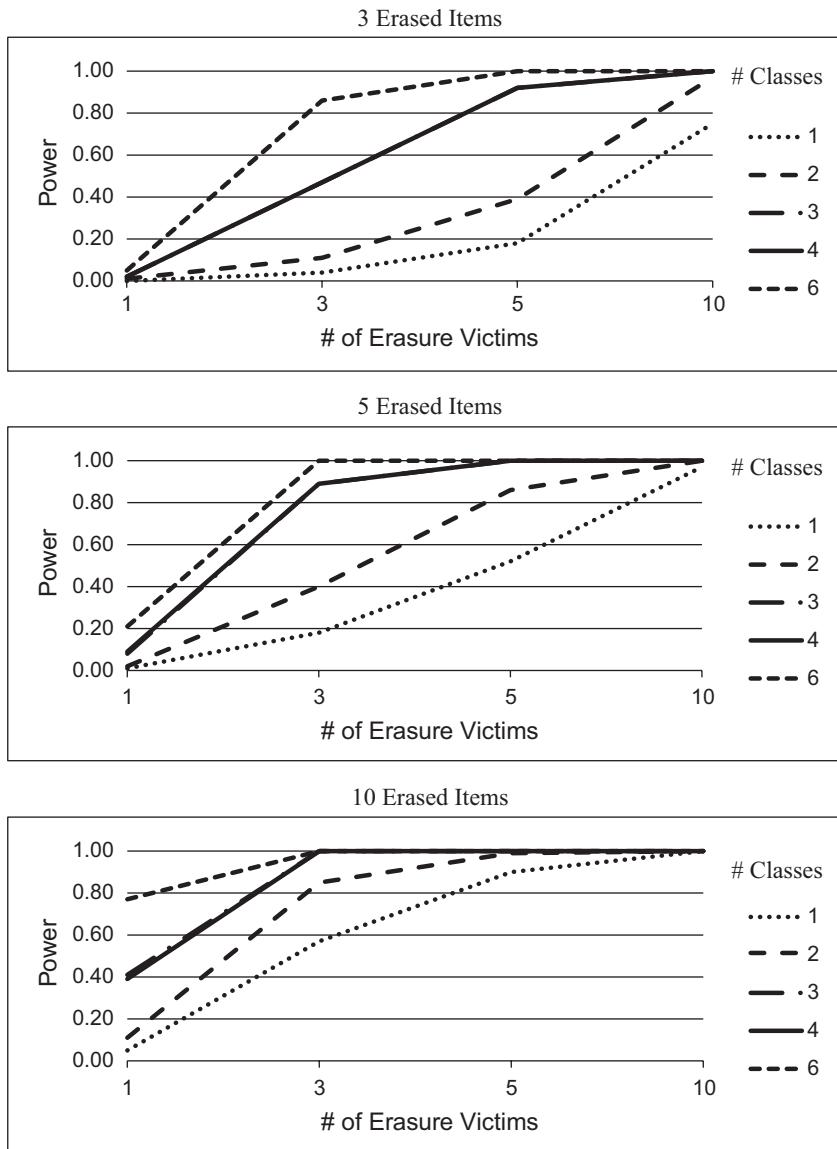
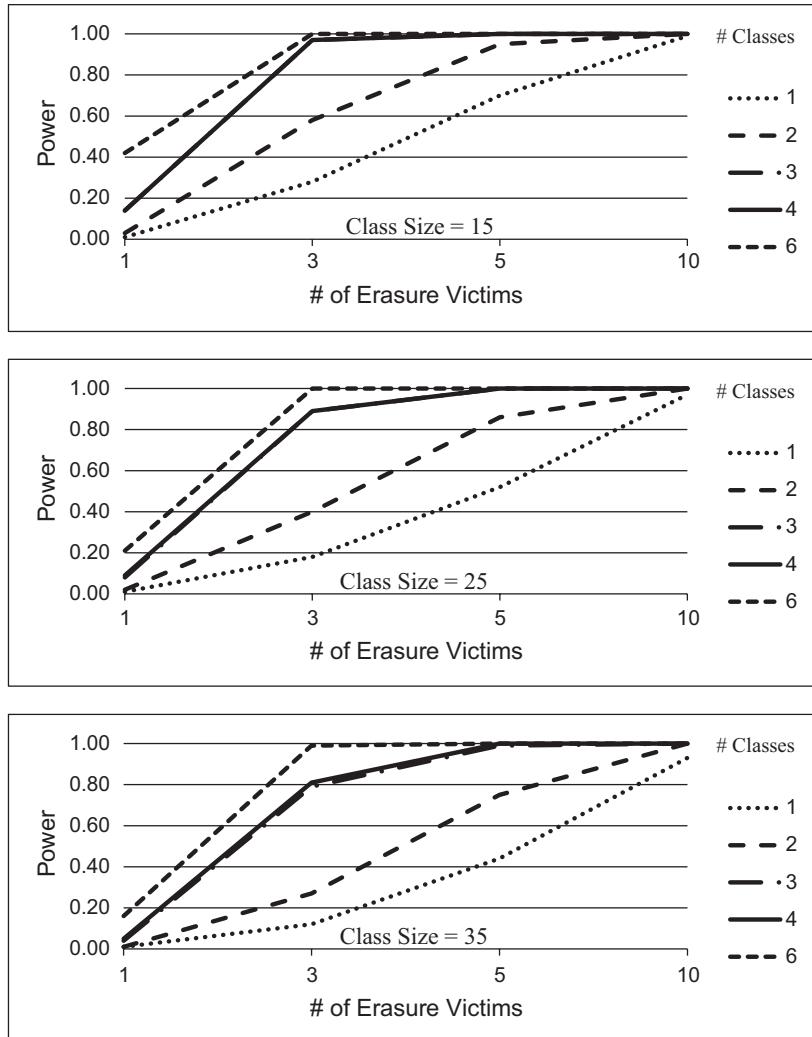


Figure 11.3 Power (at $\alpha = .001$) to Detect Schools, Based on Numbers of Tampered Classes, Erasure Victims, and Tampered Items

Note: Class size = 25.

For a fixed number of tampered classes, the general pattern of results is highly similar to those found when detecting classes. However, in detecting schools, power increases significantly as the number of tampered classes increases. The one apparent exception to this is that the results are very similar between three and four tampered classes—so similar, in fact, that the lines for the two levels are almost entirely superimposed. This anomaly is a result of the simulation design. In looking at Table 11.1, one can see that, given how data were simulated, whenever three tampered classes were simulated, it was within a three-class school in which all classes experienced tampering. However,

**Figure 11.4** Power (at $\alpha = .001$) to Detect Schools, Based on Numbers of Tampered Classes, Erasure Victims, and Class Size

Note: Number of erasures = 5.

whenever four tampered classes were simulated, it was within a six-class school in which there were also two classes which did not experience tampering and contributed null data towards the school-level statistic. Therefore, it is our expectation that had other contributing factors been equal, schools with four tampered classes would have higher power than schools with three tampered classes. Here, whatever incremental improvement would have resulted from another tampered class was entirely offset by the inclusion of data from two classes of students without tampering.

APPLICATION TO COMMON K-12 DATASET

To demonstrate the utility of EDI_g in practice, the index was applied to the common K-12 dataset to identify tampering at the class, school, and district levels. Because EDI_g evaluates tampering through erasures, rather than change scores, only one year of data

was necessary. Consequently, only fifth graders during Year 2 were used for this analysis. The resulting dataset included 72,686 students in 3,213 classes² in 1,187 schools in 630 districts. The dataset included data on 53 operational items.

An $\alpha = .001$ was used for all analyses. Although we recognize that, in practice, it is likely that states/test sponsors would apply a more conservative criterion, we felt that $\alpha = .001$ was appropriate here because (a) no investigation or potential sanctions will follow as a result of these analyses, and (b) in the interest of understanding the dataset better, we would like to detect as much tampering as possible, while still limiting the number of false positives. Given the number of classes, schools, and districts in the dataset, an $\alpha = .001$ criterion is expected to falsely identify three to four classes, one school, and zero to one districts.

The common K-12 dataset included Rasch-based item parameters, which could have been used to estimate examinee ability and determine the probability of WTR erasures. However, in the interest of consistency with the simulation study, erased items were treated as missing data, and item and ability parameters were estimated under the NRM, using MULTILOG (Thissen, 2003).

The results are shown in Table 11.7. A total of four districts, seven schools, and eight classes were identified. Each row of the table shows districts, schools, and classes that are nested. As an example, in row 4, Class 5010 was within School 244544, which was within District 71771. All significant districts also had one school identified, and in half the cases, one class was also identified. There were also several schools that were identified, even though the district was not (i.e., cases 5–7), as well as some classes that were significant even though schools were not.³ This is consistent with the results from the simulation study, which showed that for fixed amounts of tampering, smaller groups will be easier to detect than large groups.⁴ In the few instances where a larger group (e.g., School) was detected without detecting a smaller group (e.g., Class), it was always because no class was specified; hence, a class-level analysis was not performed within that school.

Table 11.7 Flagged Districts, Schools, and Classes in Common Education Dataset

	<i>Districts</i>		<i>Schools</i>		<i>Classes</i>	
	<i>ID #</i>	<i>EDI_g</i>	<i>ID #</i>	<i>EDI_g</i>	<i>ID #</i>	<i>EDI_g</i>
1.	401600	6.54	344969	6.54	9	7.76
2.	274475	4.29	273425	5.75	—	—
3.	13758	3.41	13758	5.58	—	—
4.	71771	3.37	244544	3.21	5010	3.09
5.	—	—	391665	4.40	—	—
6.	—	—	354770	3.98	331	4.11
7.	—	—	187462	3.37	—	—
8.	—	—	241507	NS	3667	4.00
					3666	3.16
9.	—	—	15517	NS	5108	3.25
10.	—	—	388551	NS	102	3.26
11.	—	—	216471	NS	1000	3.08

Note: Flagging criterion was $\alpha = .001$. School IDs are provided for all flagged classes because class IDs are unique only within School. Schools with EDI_g values below the critical value are indicated as Not Significant (NS).

The number of significant EDI_g indexes are pretty small, though clearly more than would be suggested by chance alone. Given the performance of EDI_g in the simulation, there are two possible explanations for the small number of detections. First, erasures were more common in this dataset than what has generally been found. The average number of erasures in this dataset was 2.00, meaning that, on average, students erased approximately 3.7% of the items. This is nearly double the erasure rate that is generally found in educational assessments (Bishop et al., 2011; Maynes, 2013; Mroch et al., 2012; Primoli, 2012; Primoli et al., 2011; Qualls, 2001) and is nearly twice the erasure rate that was simulated here. If these erasures were caused by tampering, we would expect that the high counts were attributable to a small number of districts, schools, or states, and that EDI_g would be highly sensitive to this information. On the other hand, if the high erasure counts were attributable to benign erasures (perhaps something about the answer sheet, test layout, question design, or test timing that led to students either miskilling their answers or reconsidering and ultimately changing their selections), this would introduce noise into the EDI_g estimation and would water down any tampering effect that was present.

Second, it is entirely possible that this dataset did not include much tampering. Although the average number of erasures was higher than typical, there did not appear to be an unusually high percentage of WTR erasures. The average number of WTR erasures was 1.02, for a rate of 1.9% of the items. Although this WTR erasure rate is higher than usually seen, given the overall erasure rate, only 50.9% of the total erasures were of the WTR variety. Furthermore, in the entire dataset, there were only 157 students (0.2%) with 10 or more WTR erasures, and only two districts had five students with 10 or more WTR erasures. Both of these districts were among the 20 largest in the dataset, and one was the single largest, containing 2,398 students. In addition, there was only one school with as many as four students with 10 WTR erasures; while this school was not detected by EDI_g , two classes within the school were identified as anomalous (see Case 8 in Table 11.7). Consequently, there is ample reason to believe that the more extreme tampering conditions simulated (e.g., 10 tampered items or 5–10 erasure victims per class) were not representative of the magnitude of tampering that may have been present in the empirical dataset.

What is particularly interesting about the results of the real data analysis is that it was very difficult to use traditional test tampering indicators to “validate” the identified groups. That is, regardless of whether the data were sorted by WTR or total erasure counts, WTR or total erasure averages, or the proportion of total erasures which were of the WTR variety (WTR%), the detected districts, schools, and classes did not rank among the very highest. In fact, the correlations between EDI_g and these other variables were not particularly strong, and for several indexes, were essentially zero. These correlations are provided in Table 11.8. Table 11.8 shows seven different indexes commonly used and regarded as reasonable indicators of tampering. The total erasures and total WTR erasures are the counts of the total number of erasures or WTR erasures in the group, respectively. Average erasures and average WTR erasures standardize the total and WTR erasure counts by considering group size. These statistics are computed in two different ways, once using the group sample size (N) and the other using the count of the number of individuals in the group with at least one erasure (N^*). WTR% is the WTR count divided by the total erasure count.

As one can see, correlations with EDI_g become stronger as the group size gets smaller, but even in the best of cases, the correlations are low. The class-level WTR% is the only

Table 11.8 Correlations Between EDI_g and Other Common Indicators of Tampering

	$EDI_{District}$	EDI_{School}	EDI_{Class}
Total Erasures	.03	.01	.05
Total WTR Erasures	.07	.10	.26
Avg. Erasures (N)	.14	.13	.07
Avg. WTR Erasures (N)	.33	.30	.38
Avg. Erasures (N*)	.15	.12	.06
Avg. WTR Erasures (N*)	.37	.36	.42
WTR%	.34	.32	.51

Note: (N) means that the average was computed across all students in the group. (N*) means that the average was computed across all students with at least one erasure.

indicator that correlates at least 0.50 with EDI_g . On the one hand, this is surprising, because the statistic of primary interest in EDI_g is the group-level WTR count. However, in computing EDI_g , WTR is not used in isolation but is corrected for the expected WTR count. In fact, within this dataset, many of the individuals with the highest WTR counts were examinees who, on the remaining items, performed very well, thereby producing very high expected WTR counts. This suggests that these examinees were, perhaps, making alignment errors. However, it is also possible that these were individuals who were made to look like good examinees by changing their responses selectively to almost all the items that they did not answer correctly. The fact that these individuals were never concentrated in the same classes/schools/districts lends some support to the hypothesis that they were, in fact, honest and benign erasures; however, a method that is sensitive to change scores would be better positioned to identify whether these were students who had scored very low on previous administrations.

The finding that correlations between EDI_g and other common indicators are so low, coupled with the findings from the simulation study here showing that EDI_g has strong power to detect moderate-to-high amounts of tampering, clearly suggests that these other methods should be studied more before continuing to use them. If they prove unpowerful, test sponsors should stop using them immediately. If some of those indicators can be shown to work well, yet still produce different results, it may be that using them in combination with EDI_g could result in greater power to detect tampering than either approach in isolation.

CONCLUSION

Test tampering is but one way that educators can cheat on tests. Other forms of cheating (or, more generally, inappropriate behavior), include activities such as teaching to the test (i.e., illegal coaching), grading constructed response sections for one's own students, checking students' work during the test for signs of carelessness or unexpected inaccuracies, and encouraging students to stay home during testing days either because they are unlikely to be successful or to keep the numbers of students in targeted subgroups below the minimum required for reporting. In the case of these behaviors, the line between acceptable and unacceptable can be gray. At what point does teaching the skills likely to be assessed on an exam (and covered by state standards) cross over to become illegal coaching? Are teachers not capable of grading their own students' accountability exams with integrity, much like they do all their classroom exams

throughout the rest of the year? If a student filled in an answer to a question too lightly or legitimately failed to erase an item well enough for the scanner to detect, wouldn't the validity of the test score be bolstered by taking steps to make sure that those items are scored correctly? Although the answers to these questions may seem straightforward to those in the measurement community, they are often less clear to those administering the tests. Consequently, it is incumbent upon the measurement community to develop materials designed to educate people about the intersection of test administration and test security, especially as it relates to K-12 testing. Several good resources have recently been released (Association of Test Publishers [ATP] & National College Testing Association, 2015; Council of Chief State School Officers and ATP, 2010, 2013; Fremer & Olson, 2015; National Council on Measurement in Education, 2012; Olson & Fremer, 2013; Wollack & Case, 2016), and the hope is that as testing practices change, these resources will continue to evolve and will become standard bookshelf items in Education Departments and school and district offices across the country.

Unlike the educator behaviors mentioned above, test tampering is not gray. Although educators who have been caught in test tampering scandals invariably cite reasons such as unreasonable external pressures or trying to help students indirectly through maintaining school staff and school resources, those involved in test tampering are fully aware that they are violating test protocol. Test tampering represents a violation of the public's trust and deprives students and parents from getting accurate information about the child's academic progress that might allow the student to get the help that they need to be more successful. Therefore, to the extent that such behavior cannot be prevented, it is critical that we have appropriate tools to detect it. The EDI_g is a major step forward in this pursuit, because not only is it the first model-based approach using erasure data that has been studied in group settings but the results of the simulation here show that it has strong Type I error control and very good power to detect moderate-to-large security breaches.

Although every year there are many newspaper clippings about different test-tampering incidents across the country, it is important to be mindful of the fact that the overwhelming majority of educators are playing by the rules and are not engaging in inappropriate behaviors. The most credible estimate of the incidence rate of test tampering among educators is between 1% and 2% (Fremer, 2011). Jacob and Levitt (2003) estimated the tampering rate at between 4% and 5%, but although this study was very well executed, it was also performed in only a single city and was conducted before methods to detect tampering had been developed, before states were routinely mining data for evidence of tampering, and before educators had lost their jobs or been issued jail sentences for changing students' answers. It is our sincere hope that as more sophisticated methods, such as EDI_g , are developed and utilized, that the incidence and magnitude of teacher cheating will decrease and that the public trust in educators and test score inferences on accountability tests will be restored.

NOTES

1. Note that Wollack et al. (2015) also explored EDI without continuity correction, but found that the Type I error rates were significantly inflated, especially for lower achieving students, the precise population for whom tampering figures to be a problem. Hence, it is being presented here only with the continuity correction.
2. There were many students who were assigned to a district and a school, but not to a class. This class total reflects only those students for whom a class code was identified. Students not issued a class code were still included in school- and district-level analyses.

3. In this case, the school ID is provided in Table 11.7 to help identify the class, because class IDs are unique only within school.
4. It is also worth noting that the expected number of false positives was largest for classes and smallest for districts.

REFERENCES

- Association of Test Publishers & National College Testing Association (2015). *Proctoring best practices*. Author.
- Bishop, S., & Egan, K. (this volume). Detecting erasures and unusual gain scores: Understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.
- Bishop, N. S., Liassou, D., Bulut, O., Seo, D. G., & Bishop, K. (2011, April). *Modeling erasure behavior*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 46, 443–459.
- Council of Chief State School Officers and Association of Test Publishers. (2010). *Operational best practices for statewide large-scale assessment programs*. Washington, DC: Author.
- Council of Chief State School Officers and Association of Test Publishers. (2013). *Operational best practices for statewide large-scale assessment programs: 2013 edition*. Washington, DC: Author.
- Fremer, J. (2011, August). *What everyone wants to know about cheating in schools*. Caveon Test Security. Retrieved from www.caveon.com/what-everyone-wants-to-know-about-cheating-in-schools/.
- Fremer, J., & Olson, J. (2015). *TILSA test security: Lessons learned by State Assessment Programs in preventing, detecting, and investigating test security irregularities*. Washington, DC: Council of Chief State School Officers.
- Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118, 843–877.
- Maynes, D. (2013). Educator cheating and the statistical detection of group-based test security threats. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 173–199). New York, NY: Routledge.
- Mroch, A. A., Lu, Y., Huang, C.-Y., & Harris, D. J. (2012, May). *Patterns of erasure behavior for a large-scale assessment*. Paper presented for the Conference on the Statistical Detection of Potential Test Fraud, Lawrence, KS.
- National Council on Measurement in Education. (2012). *Testing and data integrity in the administration of statewide student assessment programs*. Madison, WI: Author.
- Olson, J., & Fremer, J. (2013). *TILSA test security guidebook: Preventing, detecting, and investigating test security irregularities*. Washington, DC: Council of Chief State School Officers.
- Primoli, V. (2012, May). *AYP consequences and erasure behavior*. Paper presented at the annual meeting of the Conference on the Statistical Detection of Potential Test Fraud, Lawrence, KS.
- Primoli, V., Liassou, D., Bishop, N. S., & Nhousyvanisvong, A. (2011, April). *Erasure descriptive statistics and covariates*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Qualls, A. L. (2001). Can knowledge of erasure behavior be used as an indicator of possible cheating? *Educational Measurement: Issues and Practice*, 20(1), 9–16.
- Thissen, D. (2003). *MULTILOG 7.0: Multiple, categorical item analysis and test scoring using item response theory* [Computer program]. Chicago, IL: Scientific Software.
- van der Linden, W. J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics*, 37, 180–199.
- van der Linden, W. J., & Sotaridona, L. S. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, 31, 283–304.
- Wollack, J. A. & Case, S. M. (2016). Maintaining fairness through test administration. In N. J. Dorans & L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 33–53). New York, NY: Routledge.
- Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement*, 21(4), 307–320.
- Wollack, J. A., Cohen, A. S., & Eckerly, C. A. (2015). Detecting test tampering using item response theory. *Educational and Psychological Measurement*, 75, 931–953.
- Wollack, J. A. & Maynes, D. (2014, April). *Improving the robustness of erasure detection to scanner undercounting*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

12

A BAYESIAN HIERARCHICAL MODEL FOR DETECTING ABERRANT GROWTH AT THE GROUP LEVEL

William P. Skorupski, Joe Fitzpatrick, and Karla Egan

INTRODUCTION

Cheating has been a problem in testing for as long as there have been high stakes associated with the results. Cheating on statewide assessments may have serious implications for the psychometric integrity of item parameters and test scores as well as the validity of those test scores. Cheaters cheat in lots of different ways, which can make the phenomenon difficult to detect. Over the last few decades, a number of approaches for detecting cheating have been suggested, such as identifying unusual similarity among response patterns (e.g., Wollack, 1997, 2003) and analyzing person-fit data (e.g., Drasgow, Levine, & Williams, 1985; Levine & Rubin, 1979). These techniques operationalize cheating as something individual test takers do, either by copying answers or by using illicit materials to enhance their scores. The stakes in testing, however, are not always high for only the student. In some cases, teacher merit pay is tied to test results. Under adequate yearly progress (AYP) standards associated with No Child Left Behind (NCLB), schools and even districts may be shut down or taken over by the state based on test results. In K-12 statewide assessments, teachers and administrators may be motivated to cheat because test results are often used for teacher, school, and district accountability. This cheating may range from subtle (e.g., teaching to the test) to blatant (e.g., changing student answer documents). The threat of not making AYP provides a large incentive for educators to cheat. In many states, AYP is based, in part, on the percentage of students reaching proficiency. Other states have started using growth models based on individual student growth across years when estimating AYP.

Thiessen (2007) estimates that 25% of educators cheat on standardized tests in some way. Jacob and Levitt (2004) examined fluctuations in student test scores and similarity of answer patterns to detect classroom-level cheating incidents at a Chicago area school. Their work uncovered cheating incidences in 4% to 5% of classrooms studied. Recent news stories, such as the Atlanta Public Schools cheating scandal (see www.ajc.com, the website for the *Atlanta Journal-Constitution*, for detailed coverage)

highlight the fact that this phenomenon is shockingly prevalent. In Atlanta, dozens of teachers and administrators have been indicted for changing students' answer to improve scores, and many have been convicted of racketeering for involvement in these activities. These findings call into question the validity of using test scores to make important decisions.

The purpose of this chapter is to explicate a method for evaluating group-level aberrance as potential evidence of cheating. Group-level cheating refers to some kind of adult intervention with respect to students' test scores (e.g., unusual score gains for students within a classroom, answer changing). This method focuses on unusual score gains as potential evidence of cheating. Recently, Skorupski and Egan (2013, 2014) have introduced a statistical method for detecting possible group-level cheating using a Bayesian Hierarchical Linear Model (BHLM). The BHLM models the change in individual scores, nested within groups (schools) over time. After accounting for group- and time-level effects, unusual or aberrant growth is considered evidence of potential cheating. This modeling approach is both simple and generalizable. There is no need for vertically scaled assessments, and the focus is on a pair of test scores for each individual. Thus, a current year's test scores are conditioned on scores from the previous year, without extending that timeline to any additional time points. The added benefit of this approach is that students do not need to remain in intact units. In this growth model, group-level information is only incorporated at Time 2. Scores from the previous year are indexed by the individuals within that group, but they need not have been in the same classroom or school the previous year; Time 1 scores are merely the baseline of performance.

The current study further develops the method introduced in Skorupski and Egan (2013, 2014) by describing a Monte Carlo simulation study. This study evaluates the sensitivity of the method for identifying known incidents of simulated cheating and the specificity to exclude known noncheaters. Following the results of the Monte Carlo simulation study is a demonstration of the score gains analysis using the common data set from a large statewide testing program made available for this *Handbook*. The ultimate goal is to develop and validate a reliable method for identifying group-level cheating behavior such as inappropriate coaching, widespread use of improper materials during testing, or answer changing by teachers. It is hypothesized that any such group-level cheating behavior would not manifest itself using traditional individual-level cheating-detection procedures. These aberrances would likely only be detectable at a group level.

The Monte Carlo study was conducted by simulating test scores for 300 groups over two test administrations. The BHLM was fit to each replicated dataset, with estimation of a "growth aberrance" statistic used to evaluate potential cheating behavior. Groups simulated to display aberrant behavior were used for power analyses; nonaberrant groups were used to evaluate Type I error rates. An advantage of using the fully Bayesian framework was that stochastic inferences about cheating likelihood (given an aberrance criterion) could also be obtained.

METHOD

Bayesian procedures are used to estimate the parameters of a hierarchical linear model (HLM). Group-by-time means are estimated while allowing for each group to have a separately estimated variance-covariance structure across time. Growth aberrance is then inferred by means of the difference between two standardized measures of group performance. This "growth aberrance" then becomes the basis for inference making.

Simulated Data

Data were simulated for this study to emulate 2 years of student standardized test scores. To evaluate the ability of this approach to evaluate classroom-level effects, 300 groups of examinees, uniformly ranging between $n = 5$ and $n = 35$, were generated. These values were chosen so that the mean classroom size was 20, which reflects the U.S. national average (U.S. Department of Education, National Center for Education Statistics, 2012). Thus, for a given condition of the study, the mean total N was 6,000. Although these groups are referred to as “classrooms,” they could just as easily represent all but the largest of schools. Because a real-data analysis would be conducted within a grade level and subject, the within-group sample size represents all the individuals within that unit of analysis (which could be multiple classrooms within a school). However, because the analysis of statistical power is particularly important here, more groups with generally smaller sample sizes were simulated, because power for even larger groups will always increase.

Examinees within 300 groups were simulated over two time points, with a mean increase of half a standard deviation (0.5σ) from one administration to the next. These values were chosen based on what was observed in real data from Skorupski and Egan (2011) and are reasonable estimates for one year of typical achievement growth. Of the 300 groups, either 1% (three groups) or 5% (15 groups) were simulated to appear aberrant with respect to growth, using a procedure described later in this section.

Individual scores¹ at Time 1, Y_{ig1} , were simulated by sampling from a standard normal distribution:

$$Y_{ig1} \sim N(0, 1). \quad (1)$$

It should be noted that even though the scores, Y_{ig1} , are referenced by individual (i), group (g), and time (1), all examinees are randomly sampled from the same standard normal distribution. This was done to make any group differences at Time 1 completely random. Group-level information is only evaluated for growth at Time 2. Students within a group at Time 2 will have their own Time 1 scores used as a baseline for previous achievement, but Time 1 scores are not evaluated for aberrance. Thus, students do not have to have been in the same units across time points.

Individual scores at Time 2, Y_{ig2} , were simulated by introducing group-level growth, conditional on scores at Time 1, plus individual random error and a possible “cheating effect” for those classrooms simulated to be aberrant:

$$Y_{ig2} = \mu_{g2} + \beta + \rho Y_{ig1} + \varepsilon_{ig2}, \quad (2)$$

where μ_{g2} is the mean increase of scores for students within group g at Time 2, $\mu_{g2} \sim N(0.5, 0.1^2)$. Groups means are centered about zero at Time 1, and centered about 0.5 at Time 2, thus the marginal increase in scores will be equal to 0.5σ . β is the “cheating effect,” a boost in scores for students within cheating classrooms ($\beta = 0$ for nonaberrant classrooms, and $\beta = 0.5$ or 1.0 for aberrant classrooms). ρ is the correlation between scores at Time 1 and Time 2 ($\rho = 0.7$, a value chosen based on the real-data study by Skorupski & Egan, 2011). ε_{ig2} is the random individual error, $\varepsilon_{ig2} \sim N(0, 1 - \rho^2)$. Scaling scores at Time 1 and Time 2 in this way ensured that simulated samples would have approximate unit variance and share a marginal correlation equal to ρ . Because the group sizes are relatively small, sample estimates of variance and covariance could vary widely, thus suggesting the need for a HLM approach.

In all, three factors were manipulated to evaluate their impact on aberrance detection: (1) the size of the “cheating effect” (β from equation 2), (2) the type of cheating, and (3) the percentage of groups simulated to be aberrant. Each of these factors had two levels that were completely crossed. The size of the cheating effect, β , was either 0.5 or 1.0. The type of cheating was simulated in two ways: (1) the effect β was as described in Equation 2, a simple additive constant which affects only the mean of the group, or (2) the effect of β increased the mean of the group, but also decreased the within-group variance. This level was selected because it is hypothesized that teacher cheating (for example, changing students’ answers to make more of them correct) will likely result in not only artificially increasing the group’s mean but also artificially deflating the group’s variance. To simulate this second kind of cheating (henceforth referred to as “Homogenized”), scores for students within aberrant groups at Time 2 were simulated by simply sampling from a univariate normal distribution:

$$Y_{ig2} \sim N(\mu_{g2} + \beta, 0.5^2). \quad (3)$$

Thus, one would expect students in these groups to have smaller variance at Time 2 and smaller covariance between time points. This is also suggestive of the HLM approach being employed.

Each of the three factors was completely crossed, resulting in eight total conditions. Data were simulated for 300 groups with a mean class size of 20 nested individuals (therefore, the mean total sample size was 6,000) over two time points. This process was replicated 50 times per condition, resulting in 400 simulated datasets. Each of these replicated datasets was analyzed with the BHLM approach, and results were averaged over replications to insure reliability.

Analysis

The BHLM was fit to each of the 400 replicated datasets with scores over time nested within students, who in turn are nested within groups. The parameters were estimated within a fully Bayesian network implemented with Markov Chain Monte Carlo (MCMC) techniques using the freeware OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009). Sample code for running these analyses in OpenBUGS is contained in Appendix C.

The basic premise of a hierarchical model is to establish a series of nested equations, wherein independent variables from one level of the model become dependent variables at the next level. In this approach, individuals are nested within group-by-time units, and these effects are in turn nested within time points. The nested models were as follows:

$$Y_{igt} = \mu_{gt} + \varepsilon_{igt}, \text{ and} \quad (4)$$

$$\mu_{gt} = \mu_t + \tau_{gt}, \quad (5)$$

where Y_{igt} is the score of student i in group g at time t , with $i = 1, \dots, N(g)$; $N(g)$ is the number of individuals in group g ; $g = 1, \dots, 300$; and $t = 1, 2$; a simple hierarchical model with random intercepts for groups-by-times. However, an important factor in

the estimation of this model is that the individual error variance-covariance structure is estimated separately for each group:

$$\varepsilon_{igt} \sim \text{MVN}(\underline{0}, \Sigma_g), \text{ and} \quad (6)$$

$$\tau_{gt} \sim \text{MVN}(\underline{0}, \psi). \quad (7)$$

Thus, this modeling approach can accommodate different variances at each time, and different variances and covariances across time within each group. Using this approach, one can monitor individual growth of students within units over time without having to rely on a vertical scale. It should be noted that this model, as specified, only considers random intercepts for groups, which was appropriate given the simulation. However, other applications of this method could include individual, group, or time covariates, to control for growth that may be associated with them.

After fitting the model, posterior distributions were evaluated for the convergence of their solutions. The successful convergence of model parameters from the MCMC process was assessed using techniques suggested by Gelman, Carlin, Stern, and Rubin (2004). Parameter estimates from OpenBUGS were used to evaluate growth aberrance. Growth Aberrance (GA) was calculated as a difference between two effect sizes. The delta (δ) statistic (Cohen, 1988) is a simple, standardized measure of an effect size that does not incorporate sample size information. The result is that delta is equal to an estimate of the difference between a group's performance and an expected baseline level in standard deviation units. Its general form is to divide an observed difference by an estimate of the population standard deviation. To standardize the metric for growth, and obviate the need for a vertical scale, these effect sizes were calculated within each year before calculating their difference:

$$GA_g = \frac{\hat{\mu}_{g2} - \hat{\mu}_2}{\sqrt{\sigma_{g2}^2}} - \frac{\hat{\mu}_{g1} - \hat{\mu}_1}{\sqrt{\sigma_{g1}^2}}, \quad (8)$$

where μ_{gt} is the mean of group g at time t , μ_t is the marginal mean at time t , σ_{gt}^2 is the variance of scores for group g at time t (σ_{gt}^2 is the t -th diagonal element of Σ_g , the variance-covariance matrix for group g). Thus, GA_g is not a measure of growth, but a measure of growth above baseline, because the calculation accounts for the marginal differences between μ_2 and μ_1 . To be conservative with respect to Type I error, $GA_g \geq 1$ was considered a large effect and used as a criterion for flagging groups as aberrant. In practical applications, units flagged as potential cheaters would then require further review to determine whether this aberrant growth was commendable or condemnable.

The parameters of these models could be evaluated using any software capable of handling multilevel data (e.g., HLM or SEM software). However, this model was fit using OpenBUGS to take advantage of the stochastic inference MCMC output can provide. An MCMC algorithm produces random draws from the posterior distribution of each parameter being estimated. These values can be used to make any probabilistic inference without assuming a known density function for the parameter. For example, one can iteratively test a series of potential "cutscores" used for flagging potential cheating by specifying a baseline expected value and determining the posterior probability that a group's performance is unusual in comparison. This probability is easily calculated once the MCMC output is available by counting the number of random

posterior draws appearing above the cutscore and then by dividing that number by the total posterior draws available.

This approach is methodologically quite similar to a method developed by Wainer, Wang, Skorupski, and Bradlow (2005) for evaluating the reliability of pass/fail decisions. Wainer et al. define the Posterior Probability of Passing for an examinee as the proportion of posterior draws above a given cutscore. Adapted from this, the current approach is to define the Posterior Probability of Cheating (PPoC) for a group based on the proportion of GA_g posterior draws above a given threshold. For this study, a baseline was established by treating the threshold as zero so the PPoC is equal to the proportion of GA_g posterior samples greater than zero.

Cross Validation of Cheating Detection

Based on analyses in previous studies, both GA_g and PPoC were used to flag potentially cheating groups. To reduce the probability of Type I error, relatively strict criteria were employed for detection. Growth aberrance was flagged as aberrant if GA_g was greater than or equal to 1.0 and $PPoC_{gt}$ was 0.8 or greater (indicating aberrant growth at least one standard deviation above expectation with an 80% or greater chance of being a greater-than-zero effect). These values were selected based on real-data analyses from the previous study, which maintained an overall detection rate of less than 10%. It is important to note that such statistical criteria are never proof of cheating. It is even possible that these criteria could alternatively be used to identify exemplary groups. Conversely, group estimates in the extreme opposite direction (i.e., decreases relative to baseline) might be used to identify at risk groups.

Evaluation of Outcomes

The primary outcomes of interest were the sensitivity and specificity of the modeling/flagging approach described in the previous sections. Three statistical criteria were considered: (1) Type I error, (2) Power, and (3) the marginal PPoC. Type I error is evaluated as the proportion of nonaberrant groups falsely flagged as potential cheaters. Power is evaluated as the proportion of aberrant groups that were correctly flagged as potential cheaters. Last, the marginal PPoC is an outcome often overlooked in such diagnostic decision-making studies (see Skorupski & Wainer, this volume, for a discussion of the importance of this outcome). While Type I error and Power are important, they are both probability statements that are conditional on knowing the true state of a group (in this case, true aberrance or not), something that can never be known in real applications. Type I error is the probability of being flagged, given a group is not aberrant. Power is the probability of being flagged, given a group is aberrant. The marginal PPoC represents the probability that a group is aberrant, given it has been flagged. In this simulation, this is easily calculated as the proportion of flagged groups that were actually simulated to be aberrant.

RESULTS AND DISCUSSION

Based on MCMC convergence diagnostics, results indicate that the parameters of the hierarchical growth model converged to stable solutions.² Parameter recovery was very good for these analyses. True time and group means were highly correlated with their estimates over conditions and replications ($r = .94$). This parameter recovery was

consistent across all conditions. Figure 12.1 contains a graphic representation of the distribution of group means at Time 1 and Time 2. As expected, these two distributions are centered about 0 and 0.5, respectively, indicating very little bias in estimation. The slight inflation in the mean at Time 2 (0.524, as opposed to 0.5) is because simulated cheating effects actually make the true marginal mean higher than 0.5. The marginal standard deviation of group means increases because individual variance remains constant while group mean differences are being exacerbated.

Table 12.1 contains summary statistics for the evaluation of outcomes for each of the eight conditions. A few patterns are evident. First, the Type I error rate was uniformly very low across all conditions. That is, this approach shows good specificity (not flagging nonaberrant groups). Power and marginal PPoC were much more variable. For every condition, the Type I error rate was less than 0.001. As one would

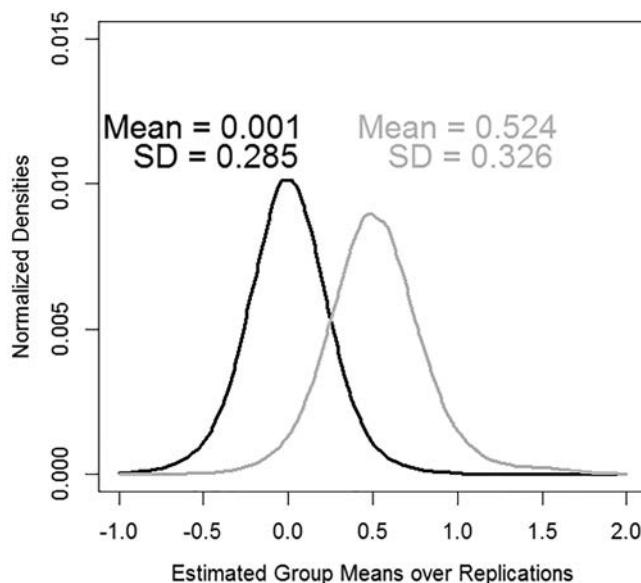


Figure 12.1 Normalized Densities of Estimated Group Means at Time 1 (black) and Time 2 (gray)

Table 12.1 Type I Error, Power, and Posterior Probability of Cheating (PPoC) for All Conditions

Magnitude of Cheating Effect	Type of Cheating	% of Groups With Simulated Cheating	Type I Error Rate	Power	Marginal PPoC
0.5	Additive	1	0.0005	0.01	0.12
0.5	Additive	5	0.0003	0.01	0.73
0.5	Homogenized	1	0.0005	0.35	0.88
0.5	Homogenized	5	0.0005	0.34	0.97
1.0	Additive	1	0.0002	0.40	0.95
1.0	Additive	5	0.0002	0.37	0.99
1.0	Homogenized	1	0.0005	0.95	0.95
1.0	Homogenized	5	0.0003	0.94	0.99

expect, Power was higher when the magnitude of the cheating effect was larger. For the $\beta = 0.5$ conditions, Power ranged between 0.01 and 0.35. Power was much better for the $\beta = 1.0$ conditions, ranging between 0.37 and 0.95. Fluctuations in Power were very clearly related to type of cheating (Additive or Homogenized). Simple Additive cheating effects had much lower Power (and therefore higher Type II error), whereas Power was always higher for the Homogenized conditions. The percentage of groups simulated with cheating (1% or 5%) appears to have had a very small effect; Power was slightly reduced as the percentage of cheaters increased.

The last column of Table 12.1 contains the results for the marginal PPoC values. In all but one condition, these results were very encouraging. For the first condition, lower magnitude, Additive cheating effect in 1% of the groups, the PPoC was only 0.12, meaning that of all the flagged groups over replications, only 12% of them were true positives. For all other conditions, however, the marginal PPoC was at least 0.73, and ranged between 0.95 and 0.99 for the $\beta = 1$ cheating magnitude conditions. PPoC tended to be higher for Homogenized versus Additive cheating, and it also increased with increased percentage of cheating groups (i.e., because more of the flagged groups were true positives when more aberrant groups were present).

These results are presented graphically in Figures 12.2 and 12.3. Figure 12.2 shows the sampling distributions of GA_g statistics for the four $\beta = 0.5$ cheating magnitude

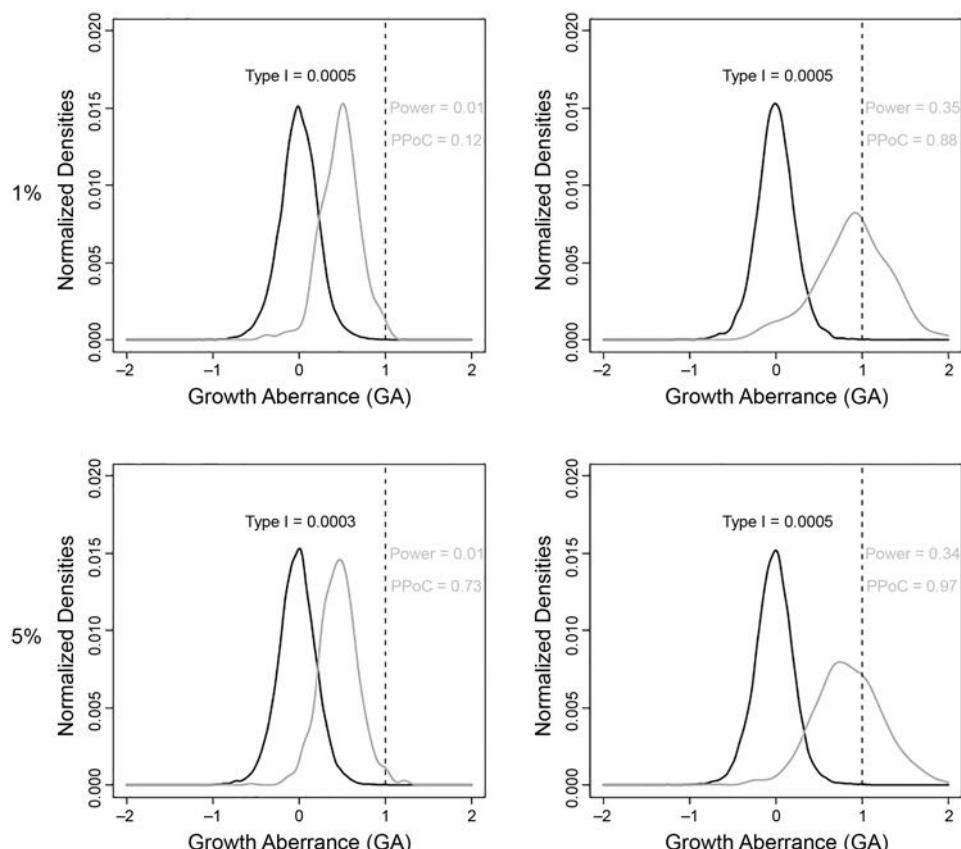


Figure 12.2 Normalized Densities of Estimated Growth Aberrance (GA) for Simulated Cheaters (gray; true Cheating Effect, $p = 0.5$) and Noncheaters (black) over replications

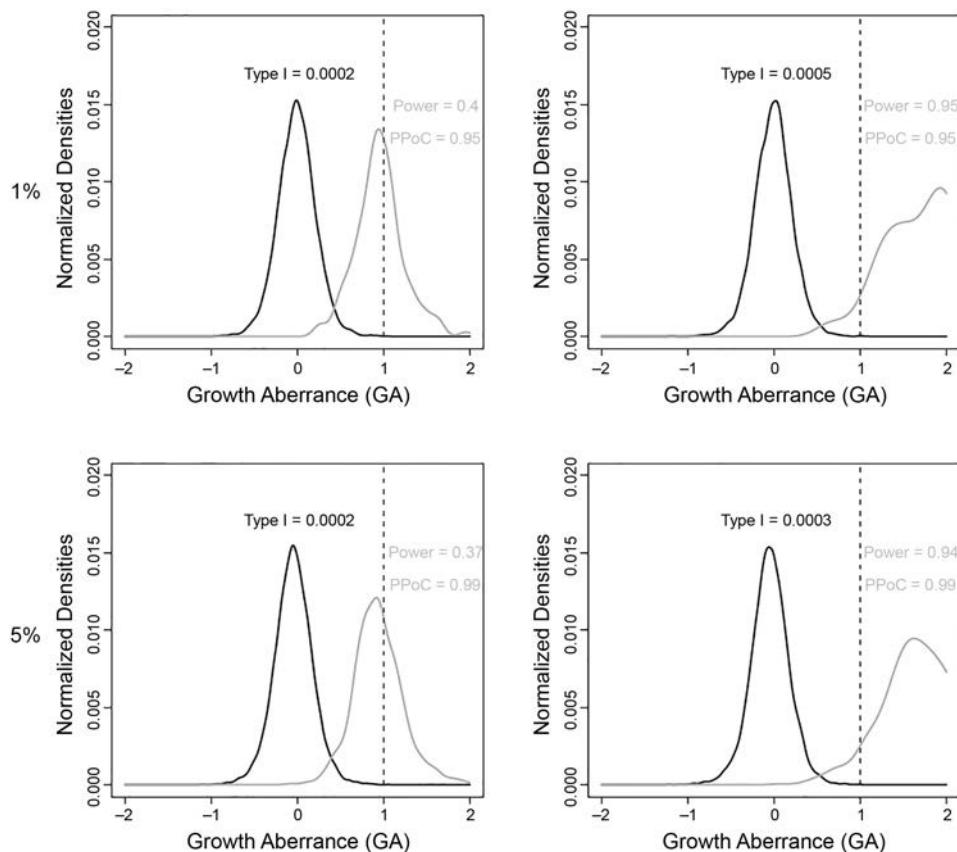


Figure 12.3 Normalized Densities of Estimated Growth Aberrance (GA) for Simulated Cheaters (gray; true Cheating Effect, $\beta = 1.0$) and Noncheaters (black) over Replications

conditions. Figure 12.3 shows the sampling distributions of GA_g statistics for the four $\beta = 1$ cheating magnitude conditions. These graphs help explain the differences in various detection rates. For the two Additive cheating conditions in Figure 12.2, one can clearly see that Power has suffered due to the use of a conservative flagging rule. However, a lower threshold for flagging would have not only increased Power but also the Type I error rate. Since the number of true negatives (nonaberrant groups) are almost always sure to vastly outnumber the true positives (aberrant groups), using a lower threshold for flagging would have reduced PPoC. For the two Homogenized cheating conditions in Figure 12.2, Power and PPoC were both improved because the same-magnitude cheating effect, coupled with reduced within-group variance, greatly benefitted sensitivity. This method will therefore likely be effective at detecting even medium aberrance if aberrant groups are also more homogenous than others.

A similar pattern, but with better detection rates, is demonstrated in Figure 12.3. The sampling distribution of GA_g is now centered about the flagging criteria for the Additive cheating effect conditions, increasing Power and dramatically increasing PPoC. For the Homogenized conditions, a large cheating magnitude coupled with reduced within-group variance for aberrant groups has resulted in extremely high sensitivity. If one assumes that a teacher purposefully manipulating test scores to improve them unfairly is likely to have a large effect on scores, then these results are very encouraging for this methodology.

DEMONSTRATION OF SCORE GAIN ANALYSIS USING REAL DATA

Description of Data

The anonymous common dataset described in this *Handbook* was used to demonstrate the BHLM. The data demonstrated here come from two years of test administration. The unit of analysis was students within schools. Only schools with five or more students in Year 2 that could be matched to students in Year 1 were retained for analysis. Based on these criteria 1,148 schools were analyzed, with a total sample size of 70,082 (86 students fewer than the original sample size).

Figure 12.4 demonstrates a histogram of the included school sample sizes. This right-skewed distribution of sample sizes shows that although many schools (219) had 30 students or fewer (and therefore most likely represented a single classroom at a grade level), a relatively small number of schools were much larger (as large as 358), representing multiple classrooms at a grade level. The BHLM approach to score gain analysis is important in such contexts because groups are likely to differ greatly by both their means and their variances; therefore, homogeneity of variance assumptions are unlikely to be met.

Analysis

A BHLM was fit to the data with scores over time nested within students, who in turn are nested within schools. The parameters were estimated within a fully Bayesian network implemented with Markov Chain Monte Carlo (MCMC) techniques using the freeware OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009), as described in the Monte Carlo simulation section. Sample code for running these analyses in OpenBUGS is contained in Appendix C. As in the simulation study, posterior distributions were evaluated for the convergence of solutions after fitting the models. Parameter estimates from OpenBUGS were used to evaluate Growth Aberrance (GA), the difference between two effect sizes. To be relatively conservative with respect to Type I error, $GA_g \geq 1.0$ with $PPoC \geq 0.9$ were the flagging criteria for aberrance (meaning that a group's

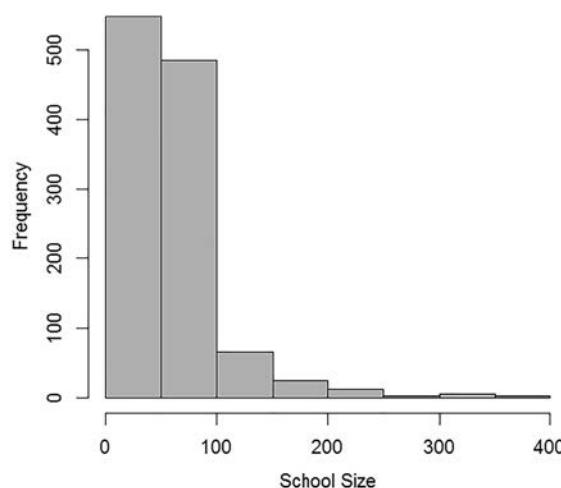


Figure 12.4 Histogram of School Sample Sizes

growth across 2 years was at least a standard deviation larger than expected, with at least a 90% chance that this effect was greater than zero).

RESULTS FOR REAL DATA ANALYSES

Based on MCMC convergence diagnostics, results indicate that the parameters of the hierarchical growth model converged to stable solutions.³ Results are presented in Figures 12.5 and 12.6 to demonstrate the amount of growth aberrance detected. Figure 12.5 contains a scatterplot of estimated school means at Year 1 and Year 2. There is

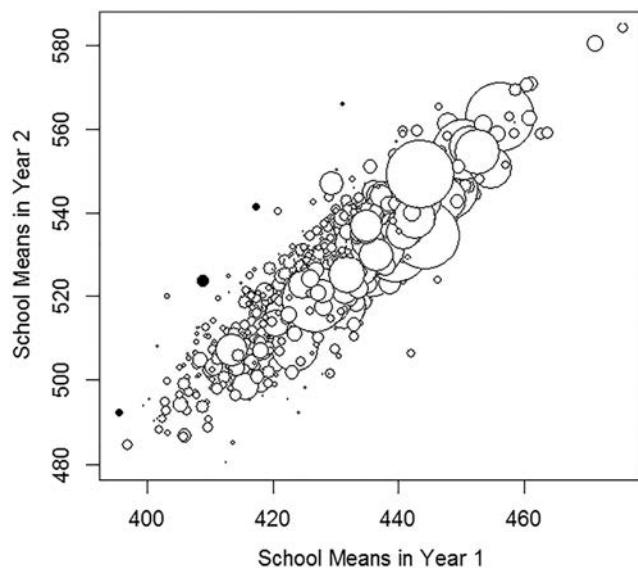


Figure 12.5 Scatterplot of School Means at Year 1 and Year 2

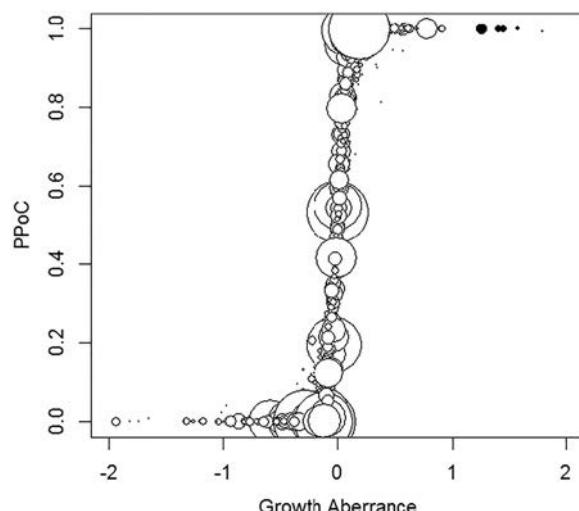


Figure 12.6 Relationship between Growth Aberrance and PPoC for Mathematics

a strong linear relationship between mean scores across years, which is not surprising. Those schools with students who demonstrated unusual growth are flagged in black in the figure, and school sizes are represented by the relative size of the circles. This plot demonstrates that the flagged schools are those with the largest, most stable positive residuals from the linear model.

Figure 12.6 demonstrates the relationship between the GA statistics and the associated PPoC values for Mathematics. Of the 1,148 schools included in this analysis, five (0.4%) were flagged as demonstrating aberrant growth (indicated by solid black). Those schools flagged are those with growth aberrances greater than 1.0 and PPoC values greater than 0.9. From this plot it is clear that one could manipulate these flagging cutoffs if a marginal detection rate of 0.004 was considered too high or too low.

CONCLUSION

Overall, this BHLM methodology appears to have great promise for detecting groups of individuals demonstrating aberrant behavior. The concept of “growth aberrance” as a statistical indicator of potential cheating provides a straightforward framework to conceptualize group-level effects. This procedure is especially effective with the BHLM, because the variance-covariance matrix is estimated separately for individual groups. This allows aberrance to be a function of unusual growth and unusual changes in variability. The PPoC provides additional insight into the probability of this aberrance being a nonzero effect. Using some reasonable criteria, this study has demonstrated that the combination of GA and PPoC values (which could be combined with any other additional available evidence, such as erasures) seems to be an effective way of achieving excellent sensitivity and specificity. Furthermore, this methodology could also be extended down to the individual level of the model, to standardize individual growth to evaluate aberrant changes for examinees. Individual effects were not simulated or considered in this study, but that could be a consideration for future research with this method.

In the simulation study, the evidence showed that the marginal PPoC was very high for most conditions, except for the hardest-to-detect first condition. PPoC was higher for the larger cheating magnitude conditions, for Homogenized versus Additive, and increased somewhat as the percentage of aberrant groups increased. Although power was not always high, PPoC results were compelling. These results are arguably the most important, because they answer the most practical question: if you used this method to flag aberrant schools as potential cheaters (as demonstrated in the real-data analyses), how often would you identify “true aberrance?” For the larger magnitude cheating effect conditions, the answer is that if a group was flagged, there was at least a 95% chance that it was an aberrant group. In general, these results support what may be a common policy preference when it comes to detecting cheating—preferring to make a Type II error (failing to identify a true cheater) than to make a Type I error (accusing an innocent of wrongdoing). The method demonstrated here adheres to that principle for all but the most subtle of cheating scenarios.

For any analysis like this one, a statistical procedure can only produce a flag that *implies* aberrant performance; it can never *prove* that, for example, teachers in a school are cheating. Thus, this procedure, like any other, will still need to be validated by gathering additional evidence and conducting investigation of those schools or classrooms that are flagged. The BHLM approach should be useful for providing more insight into the groups for which such investigations are warranted.

NOTES

1. These scores were treated as observed scores in the Bayesian HLM (that is, they were treated as being perfectly reliable). This was done to present a “best case scenario” for the approach. However, a measurement model could easily be nested within this methodology if it were desirable to account for unreliability in the estimation.
2. Parameters for each of the 400 replicated datasets were estimated by creating two independent Markov Chains for every parameter, each of which was 8,000 iterations long, with a burn-in of 6,000. Retained posterior draws from these chains all represented converged solutions.
3. Parameters for each of the two datasets were estimated by creating two independent Markov Chains for every parameter, each of which was 8,000 iterations long, with a burn-in of 6,000. Retained posterior draws from these chains all represented converged solutions.

REFERENCES

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67–86. doi:10.1111/j.2044-8317.1985.tb00817.x
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Jacob, B. A. & Levitt, S. D. (2004). To catch a cheat. *Education Next*, 4(1), 68–75.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4, 269–290. doi:10.3102/10769986004004269
- Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine*, 28, 3049–3067.
- Skorupski, W. P., & Egan, K. (2011, April). *Detecting cheating through the use of hierarchical growth models*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Skorupski, W. P., & Egan, K. (2013, October). *A Bayesian hierarchical linear model for detecting group-level cheating and aberrance*. Paper presented at the Annual Conference for the Statistical Detection of Potential Test Fraud, Madison, WI.
- Skorupski, W. P., & Egan, K. (2014). A Bayesian hierarchical linear modeling approach for detecting cheating and aberrance. In N. M. Kingston & A. Clark (eds.) *Test fraud: Statistical detection and methodology*. New York: Routledge.
- Thiessen, B. (2007). *Case study—Policies to address educator cheating*. Retrieved from: www.bradthiessen.com/html5/docs/format.pdf
- U.S. Department of Education, National Center for Education Statistics. (2012). *Digest of education statistics, 2011* (NCES 2012-001).
- Wainer, H., Wang, X., Skorupski, W. P., & Bradlow, E. T. (2005). A Bayesian method for evaluating passing scores: The PPoP curve. *Journal of Educational Measurement*, 42, 271–282. doi:10.1111/j.1745-3984.2005.00014.x
- Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement*, 21, 307–320. doi:10.1177/01466216970214002
- Wollack, J. A. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement*, 40, 189–205. doi:10.1111/j.1745-3984.2003.tb01104.x

13

USING NONLINEAR REGRESSION TO IDENTIFY UNUSUAL PERFORMANCE LEVEL CLASSIFICATION RATES

J. Michael Clark, William P. Skorupski, and Stephen Murphy

INTRODUCTION

Large-scale educational testing is a high-stakes endeavor, not just for students who may need to pass a standardized test to be promoted to the next grade level or graduate, but also for educators who may experience pressures brought on by accountability expectations related to their students' performance on these assessments. Given the ubiquity of large-scale assessments in the current K-12 educational landscape in the U.S. and the accountability pressures placed upon stakeholders, any number of individuals playing various roles in the testing process—such as test takers, teachers, proctors, educational administrators, or others—may potentially feel pressures to engage in test misconduct. Recognizing that obtaining valid test scores is a central concern for any testing program and that test misconduct represents a serious threat to test score validity (Amrein-Beardsley, Berliner, & Rideau, 2010; Cizek, 1999), it is in the best interest of educational agencies as well as their test vendors to remain vigilant for evidence of possible misconduct.

With accountability measures directly tied to student performance on state assessments, educators wishing to gain rewards or perhaps avoid sanctions might be incentivized to engage in misconduct in an effort to boost students' scores (e.g., Molland, 2012; National Center for Fair and Open Testing, 2013; Vogell, 2011). Educators who engage in misconduct may manipulate test scores in a number of possible ways, including but not limited to obtaining test items prior to the administration and sharing answers with students; allowing assistive materials such as notes, posters, or textbooks to remain accessible to students during the test; coaching or assisting students during the test administration; or erasing incorrect submitted responses and replacing them with correct answers (Cizek, 2001). When misconduct occurs, it may leave tell-tale signatures in test data, and its manifestations will likely vary depending on the form of misconduct as well as the role players responsible. If it is reasonable to assume that whichever of the aforementioned forms of misconduct results in an increase in

test-takers' observed performance on the assessment beyond what might otherwise be expected given their true levels of achievement, one straightforward approach to identify potential incidents of misconduct may be to model student achievement longitudinally, looking for greater-than-expected increases in test-taker performance at a particular time point of interest (e.g., Jacob & Levitt, 2003; NCME, 2012).

When choosing to embark upon an effort to model longitudinal changes in test performance with the explicit goal of identifying unusual changes in performance over time, it is essential that the researcher give careful thought toward key considerations in selecting the right model for the job. Of course, there is no such thing as a right or correct model in the objective sense (Box, 1976). However, some models naturally align better than others to the researcher's fundamental question, to the data, or to both—and that provides the context of the present discussion.

In the case of suspected test misconduct, one must consider several aspects of the investigation when making model design decisions. To begin, what is the unit of analysis in the investigation? At what level or levels should evidence be gathered and evaluated—for example, at the level of the examinee or at some aggregate level, such as the classroom, school, and/or district? What is the outcome measure of interest? Should predictions focus on scores or performance levels, and what might be the noteworthy trade-offs of these options? Should changes be modeled between or within cohorts of students? In other words, if the goal is to model changes in performance for a given classroom, is it most appropriate to compare how this year's class performed to how last year's class performed, or would it be better to compare observed performance for this year's class to what was predicted based on their prior test scores? Finally, how should the cutoff that defines how much growth is "too much" when making flagging decisions be defined?

As will be seen in this chapter, these are critical considerations when choosing a modeling technique. There are not often right or wrong answers to each of these important questions. As with many questions posed in the face of ambiguous situations, the best answer often depends on the question the researcher seeks to answer and the circumstances surrounding that situation. Although a particular method is described in this chapter, attention will be given to critical considerations that may lead the researcher to seek out other methods that better align with his or her research question and available data.

MODEL SELECTION CONSIDERATIONS

Assuming that (1) cheating represents a threat to the validity of test outcomes, (2) pressures may incentivize individuals to engage in misconduct, and (3) such misconduct is likely to result in artificially inflated test scores, then it stands to reason that such manipulated scores will stand out in terms of year-to-year gains, but the matter of drawing a line to identify how much of a gain is too much to be considered reasonable remains. To identify larger-than-expected gains in performance, longitudinal prediction models may be of use. However, such models offer numerous choices to consider, all of which are likely to impact how their results are used and interpreted to make decisions. A few key considerations follow.

Unit of Analysis

A primary issue the researcher must consider early in any investigation is the unit of analysis: what class of individuals is the focal unit of analysis? More specifically,

is it the students, educators, or both? A good start when framing this question is to look at pressures and incentives. Is either group pressured or incentivized to cheat? If class grades, grade-level progression, or graduation requirements are directly tied to performance on the assessment, then there may be reason to suspect that students have some incentive to cheat. Conversely, educators may have incentive to cheat to boost their students' scores due to accountability and teacher performance-pressure sources. The method described in this chapter is primarily focused on gathering evidence of systemic misconduct by educators at the classroom, school, or district level.

Choosing the Outcome Measure of Interest

To measure changes in test performance over time, the researcher must decide how such changes will be quantified. The two most salient choices are test scores and performance levels. Upon initial consideration of this dilemma, a seasoned practitioner of regression techniques will most likely lean heavily toward using test scores and shun performance levels. After all, test scores for most commonly used large-scale educational assessments provide interval-level data with a wide range of potential values, whereas performance levels are ordinal in nature and assume few values—in the most extreme case, only two (e.g., Pass/Fail). If variance truly is the lifeblood of predictive models, then it is easy to see the benefits of using test scores as the outcome measure of interest in any predictive model. Although once common, techniques for artificially dichotomizing continuous variables are now derided in the literature as a fool's errand (MacCallum, Zhang, Preacher, & Rucker, 2002). However, this chapter will make a case for consideration of performance levels instead of test scores as the outcome measure of interest based on several key considerations.

One critical difference between performance level categories obtained from an assessment and artificially dichotomized values—such as those obtained from performing a median split—is the interpretive context placed on the former. When standard setting and all associated prerequisite steps are properly executed, the resulting cuts can be used to group test scores into substantively meaningful, well-defined performance level categories. Although even under the best circumstances, the process of standard-setting necessarily involves some degree of arbitrary decision making at various points, appropriately derived performance level categories are far more meaningful in terms of interpretation than are the categories that result from purely arbitrary methods, such as median splits. Meaningfulness is critical here. Performance level classifications are central to the foundational purposes of assessment.

Consider, for example, three students who take an end-of-grade Reading test that requires students to score into the Proficient performance level to progress to the next grade. The scale cut score for this hypothetical test is 500. Student A scored at the cut, 500, and is therefore classified as Proficient. Student B scored above the cut at 510 and is also classified as Proficient. Student C scored just below the cut at 495 and has therefore failed to reach the Proficient performance level. Focusing attention on Student A (while also assuming an interval-level score scale), it is obvious that Student C's performance was closer to Student A's than to Student B's in terms of score differences, but in terms of real-world impact, the differences between scores is not nearly as impactful as is the differences in performance level outcomes. At the student level, both Students A and B have demonstrated that they meet necessary reading proficiency requirements to be promoted to the next grade, while Student C has not yet demonstrated reading

proficiency and may be looking toward a future that includes remediation, retesting, and possible retention, depending on local policies.

At the classroom, school, and district levels, performance level categories are extremely important for performance monitoring and accountability purposes. Counts of students scoring into each performance level are intensely scrutinized when issuing progress reports. For the three hypothetical students described above, the fact that two of them scored into the Proficient category and one scored below Proficient is of much greater substantive interest than the actual values of their scores. Although scale scores offer greater potential variability as an outcome measure, the researcher should always consider the real-world measurement context when choosing an appropriate modeling framework. If a five-point score difference straddling two adjacent performance levels trumps a 10-point score difference within a given performance level, then it might be reasonable to conclude that the performance level outcome is the most important outcome, and the model used to identify unusual patterns of performance changes ought to reflect this in terms of the chosen outcome variable.

Accounting for Uncertainty in Predicted Outcomes

It is impossible to overstate the importance of measurement error in the context of test theory. As long as imperfect human beings use imperfect tests to make inferences about other imperfect humans' unobservable traits, measurement error will remain an inescapable reality of the process. Through various strategies, testing professionals can seek to minimize—but never hope to fully eliminate—error in measurement. Given that measurement error is an unavoidable reality of testing, and given the potentially life-altering negative consequences of false identifications for investigations of suspected misconduct, it seems obvious that any method used to identify unusual changes in performance should properly account for potential classification errors.

In revisiting the issue of choosing the right outcome measure, it should be noted that linear models predicting scores on this year's test from prior performance will yield a single point estimate as the predicted value. If the ultimate goal is to identify expected counts of students at each performance level, one could argue that simply comparing those predicted scores to the cut scores and identifying each student's expected performance level is the most straightforward solution. On the surface, this solution would be quite easy to implement, and for many students such a practice may work satisfactorily. For some students, there may be a clear and obvious best choice when identifying the most likely performance level classification in the current year based on prior performance, particularly for those students consistently scoring at the extreme high or low ends of the achievement spectrum. However, for many examinees in the population, the picture may not be so clear.

As an example, imagine a hypothetical assessment with a Proficient scale cut of 500 and a student with a predicted score of 499.5. What would be the most appropriate classification for this student's predicted performance level? Assuming a score scale made up exclusively of whole numbers, should a predicted score of 499.5 round up to the nearest integer—therefore indicating that this student is predicted to be Proficient—or instead would it be more sensible to require that any value less than 500 (no matter how minutely so) be predicted as below Proficient? And regardless of which choice is made in this conundrum, how much confidence can reasonably be placed in the predicted performance level classification? Would either outcome—Proficient or below Proficient—be all that surprising? And finally, is it fair to the educator whose conduct

is being investigated to use such an absolute measure to quantify the expected outcomes? Should a black-or-white predicted outcome be used in a situation that seems much better served by shades of gray?

There certainly are things that could be done about this linear prediction situation. Error bands can be placed around predicted scores to identify a likely range of values. Making certain assumptions about the distributions of error around predicted scores, probabilities can be computed for various possible scores within a range surrounding the predicted score. However, the process of computing those probabilities and aggregating them into meaningful classroom-level expected counts of students for the purposes of comparison to observed counts is neither direct nor intuitive. In contrast, a multinomial model predicting performance level categories from prior scores directly aligns to the research question at hand, is easy to implement, and properly accounts for uncertainty in the predicted outcome. Revisiting the prior example, probabilities of the student scoring into each performance level are estimated from the model. So instead of obtaining a single predicted score of 499.5, this hypothetical student may be found to have a 0.501 probability of scoring below Proficient and a 0.499 probability of scoring into the Proficient category. This prediction that was before an all-or-nothing classification decision in the linear model is now a series of probabilistic values that appropriately account for uncertainty in the predicted performance level classification. Furthermore, because each student's individual probabilities sum to 1, expected counts for each performance level at the classroom, school, or district level can be computed easily by simply summing students' probabilities. For a given classroom, for example, the sum of students' individual probabilities for Performance Level 1 results in the expected number of students at this performance level and the sum of students' individual probabilities at Performance Level 2 provides the expected count at this performance level. Finally, use of this approach does not limit obtaining model-estimated probabilities to two performance levels as in the example presented here; if there are other additional performance levels, model-estimated probabilities for all performance level categories are obtained.

Cohort Modeling Approaches and Implications

It is evident that to model longitudinal changes in performance, there must be multiple measurements taken over two or more occasions, but it is worthwhile to consider the options for making these measurements. It is also instructive to consider the wider implications associated with these modeling choices. Two longitudinal measurement designs will be considered, which will be characterized in this chapter as between-cohort and within-cohort modeling options.

To employ a *between-cohort modeling* strategy, performance changes over time are modeled for a fixed unit of analysis—for example, a classroom. For such a classroom, the research question is, “How does the performance of this year’s class compare to the performance of last year’s class?” Of course, there are a number of ways that this question can be approached in a more sophisticated manner; for example, by comparing this classroom’s longitudinal change to other classrooms in the state, perhaps while controlling for covariates of interest. However, the important point here is that in a between-cohort design, the performance of the current year’s cohort of students is compared to the performance of last year’s cohort. Change is modeled between years and between students. In contrast to this approach, *within-cohort modeling* reframes the focus of the analysis to the student level. In this design, the performance of students in

a given classroom is compared to what was predicted for them based on their previous observed scores. In this context, change is modeled between years but within students.

So with two modeling approaches to consider, which option is better? It will be demonstrated neither option is necessarily superior in all situations, but it is important that the researcher tasked with investigating for evidence of possible misconduct understand the implications of this decision so that he or she can appropriately interpret outcomes. What consequences must the researcher consider when choosing how to model change? It all depends on expectations regarding cheating behavior, the nature of performance changes, and the nature of the research question.

It is important to recognize that in order for cheating to be detectable, there has to be a measurable deviation from some predefined baseline of expectation. The context in which that expectation is constructed will directly impact the researcher's ability to identify unusual patterns of test data as well as how he or she interprets outcomes. When choosing whether to model change within or between cohorts of examinees, the researcher must consider how different potential design choices might impact how an observed change in performance across time points could be interpreted. For example, suppose that for one particular classroom, student response forms were collected and altered to replace a large number of incorrect responses with correct ones. This action had the impact of raising scores for a substantial number of students in this classroom, thus boosting some students' scores above the Proficient cutoff. If the researcher chooses a between-cohort modeling strategy—which compares the students in this year's classroom to the performance of last year's students—the researcher might find success in detecting this anomaly if this is the first year in which cheating has occurred. If cheating has been going on in this classroom for many years, the researcher might reach the (incorrect) conclusion that this is a high-achieving class and find no reason for suspicion of these scores. If a within-cohort modeling design is used, however, the chances of identifying unusual score change patterns may increase. For this given classroom, suppose that instead of comparing the average score of this year's class to the average score obtained from last year's class, the researcher uses a longitudinal model to predict scores for this year's group of students from *those students' prior test scores* and aggregate those predictions in such a way that allows the researcher to construct expected outcomes, which can be compared to observed outcomes. In this case, the technique's detection power is no longer dependent on the teacher in question having just begun cheating that very year. Of course, actions taken by other teachers in prior years will necessarily impact detection outcomes; thus, in instances of widespread, systemic cheating that impacts students across grade levels and years, within-cohort modeling will likely fare no better than between-cohort modeling. However, a within-cohort modeling strategy should perform no worse than a between-cohort strategy in such a situation, either.

USING THE CUMULATIVE LOGIT MODEL TO IDENTIFY UNUSUAL PATTERNS

The modeling strategy proposed in this chapter uses aggregated results of an examinee-level nonlinear regression model to make unit-level inferences regarding the reasonableness of observed proportions of test takers classified into performance level categories (denoted $j = 1, 2, \dots, J$), conditional on students' prior scale scores. The proposed methodology models test performance longitudinally across time points and within test takers. The outcome variable in this prediction model is the student's observed performance

level category (Y), which is treated as an ordinal variable, at time t , and the predictor variable is the student's scale score obtained in the previous grade level at time $t - 1$. Therefore, this is characterized as a cumulative logit regression model (Agresti, 1996). Because students are not necessarily expected to be placed in identical classrooms—in terms of classrooms being comprised of identical peer groups—across years, students are not expected to be nested within units across time points.

As described by Agresti (1996), cumulative probabilities reflect the ordering that $P(Y \leq 1) \leq P(Y \leq 2) \leq \dots P(Y \leq J) = 1$; the cumulative probability that $Y \leq J$ necessarily equals 1, and the logits for the first $J - 1$ cumulative probabilities are

$$\text{logit}[P(Y \leq j)] = \log \left(\frac{P(Y \leq j)}{1 - P(Y \leq j)} \right) \quad (1)$$

$$= \log \left(\frac{\pi_1 + \dots + \pi_j}{\pi_{j+1} + \dots + \pi_J} \right), \quad j = 1, \dots, J - 1. \quad (2)$$

Cumulative probabilities obtained from this model are used to compute individual probabilities. For performance level category $j = 1$, the individual probability is equal to the cumulative probability for $j = 1$. For performance levels $j = 2$ through J , the individual probability is equal to the difference between the cumulative probability for category j and the cumulative probability for category $j - 1$. An example showing individual probabilities for a test with $J = 4$ performance level categories at time t is provided in Figure 13.1.

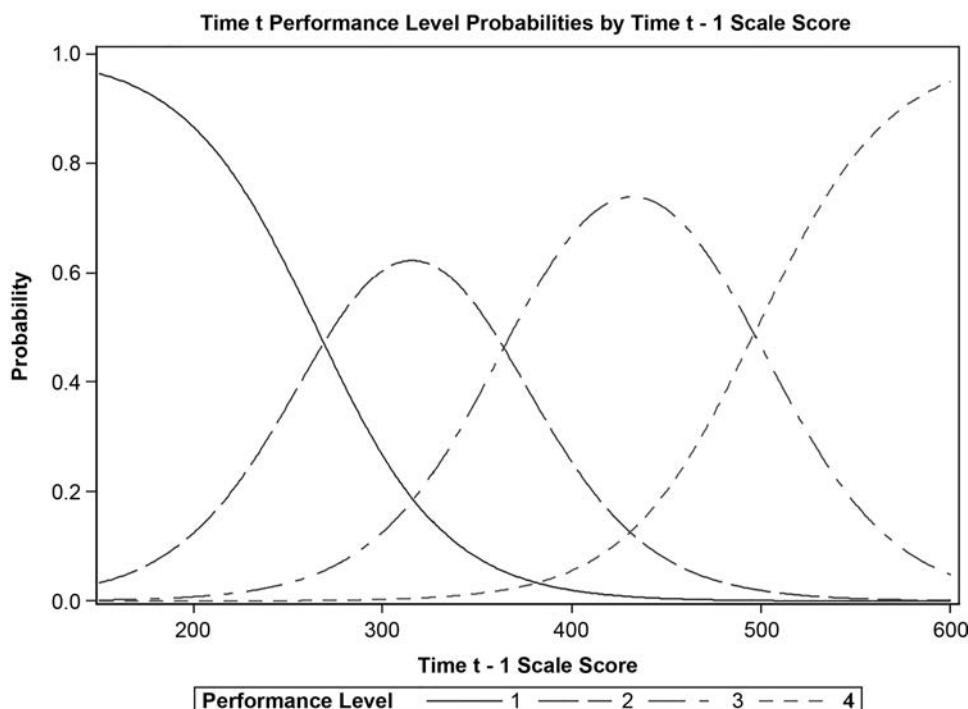


Figure 13.1 Example of Individual Probabilities for Four Performance Level Categories

As is evident from this example, independent probabilities of scoring into each of the four performance levels at time t conditional on any observed scale score from time $t - 1$ can be computed, and independent probabilities sum to 1. For example, a test taker who achieved a score of 300 at time $t - 1$ has a probability of 0.266 of scoring into Performance Level 1 at time t , a 0.605 probability of scoring into Performance Level 2, a 0.126 probability of scoring into Performance Level 3, and a 0.003 probability of scoring into Performance Level 4. For this student, it is reasonable to expect that he or she would score into Performance Level 2 at time t , although a score placing the examinee into Performance Level 1 would not be wholly unreasonable, either. A score placing this student into Performance Level 4 at time t , however, would be highly improbable according to this model.

As previously discussed, a possible alternative approach based on a linear prediction model might be useful to predict examinees' scale scores at time t from their time $t - 1$ scale scores, and then apply the performance level cuts to the predicted and observed time t scale scores to obtain predicted and observed performance level outcomes. However, as illustrated in Figure 13.1, there are regions of scale scores where multiple performance level outcomes are nearly equally probable. For example, a test taker who obtained a scale score of 364 at time $t - 1$ has a 0.461 probability of scoring into Performance Level 2 at time t , and a 0.465 probability of scoring into Performance Level 3. Although this test taker has a slightly higher probability of scoring into Performance Level 3, practically speaking, the difference is trivial and predicting either outcome would be reasonable. When aggregating expected values across students within a unit, unit-level expected values computed through this cumulative logit model will reflect this inherent uncertainty in student-level performance level classifications, particularly when multiple performance level outcomes are almost equally likely.

Computing the Standardized Residual

Treating independent probabilities as examinee-level expected values, these can be aggregated easily to compute the expected count of examinees at performance level j for unit k by summing independent probabilities for performance level j across all examinees in the unit ($i = 1, 2, \dots, n_k$), conditioning on their respective scale scores from the prior time point. Equation 3 shows that dividing these expected counts by the total number of examinees in the unit yields the expected proportion of examinees at performance level j for unit k :

$$E(P_{jk}) = \frac{\sum_{i=1}^{n_k} P_{ij}}{n_k}. \quad (3)$$

The standardized residual for performance level j for unit k is the difference between the observed proportion, P_{jk} , and the expected proportion, $E(P_{jk})$, divided by the standard error:

$$SR_{jk} = \frac{P_{jk} - E(P_{jk})}{SE_{jk}}, \quad (4)$$

where the standard error is equal to the square root of the estimated variance of the mean:

$$SE_{jk} = \sqrt{\frac{E(P_{jk})(1 - E(P_{jk}))}{n_k}}. \quad (5)$$

Standardized residuals are expected to be normally distributed with a mean of 0 and a standard deviation of 1. For any performance level, positive residuals are taken as indication of a higher-than-predicted proportion of students within the unit scoring into that performance level, and a negative residual is indication of a lower-than-predicted proportion. If, for example, a given assessment program has four performance level categories, and Performance Level 3 corresponds to Proficient or on-grade-level performance and Performance Level 4 corresponds to superior performance, then extremely large, positive residuals for either or both of these categories may be taken as a potential reason for further follow-up. In researching this methodology, the authors have generally focused exclusively on flagging large positive residuals for performance levels at or above the Proficient cut, namely due to the dependence across performance levels. That is, if a given classroom has many more students than predicted above the Proficient cutoff, then necessarily there will be fewer students than predicted below the cutoff.

Yet another benefit of any regression-based modeling strategy, including the one presented in this chapter, is that such methods are not scale dependent—meaning that it is not necessary for the scores across grade levels to be placed onto a vertical scale in order for the method to work. At no point are raw differences computed. As the prediction relies on the nonlinear relationship between prior test scores and the current year's performance levels, scaling is of no concern when constructing this model. A strong correlation in testing outcomes across time points is critical here, but the actual scaling of the assessments is not. This affords a great deal of flexibility in specifying prediction models.

Considerations for the Possible Inclusion of Student Demographic Covariates

Thus far, discussion has centered on the essential building blocks of a prediction model: individual students' performance level classifications in the current year predicted from their prior scores, with aggregated individual probabilities for each performance level providing the foundation for expected counts and proportions. Looking beyond students' prior achievement, is there any reason to statistically control for other student demographic characteristics, such as gender, race, or socioeconomic status, to name a few? If any group differences along these lines are found to exist, is it appropriate to adjust for them, and what might the potential complications be of the possible decisions to be made in this regard?

At this point, a reader with some knowledge of growth models or value-added models has already noticed parallels between the methodologies used in those pursuits and the one presented in this chapter, albeit framed in very different contexts. Whether it is necessary to statistically control for group differences attributed to student demographic variables has been thoroughly discussed in the considerable literature for growth and value-added models and will not be dealt with in detail here. However, to supplement those discussions in the present context of predictive modeling for the purposes of misconduct investigations, it is critical to consider the real-world implications of this decision.

It is no exaggeration to say that large-scale educational assessment has become a lightning rod for controversy in recent years. Whether it be due to discontent over educational policy, mistakes in the development and delivery of high-profile standardized assessments, performance pressure placed on students and educators, or any combination of these and other controversial topics surrounding K-12 assessment, intense

scrutiny is placed on the actions of state and local education agencies and their test vendors. In sensitive contexts such as data forensic investigations into possible misconduct, researchers must weigh the pros and cons of building the best possible model against likely real-world consequences, and these two pursuits are sometimes at odds with one another. In the case of misconduct investigations, a researcher must choose whether it is best to directly control for demographic characteristics of students in the prediction model or instead predict this year's outcomes based solely on prior student achievement. Unfortunately, both choices invite the potential for controversy. If the researcher elects to include only prior achievement as a predictor in the model, then this implies that all students within the population are expected to follow the same growth trajectory over time. If there are significant differences across student demographic groups in longitudinal growth patterns, fairness concerns might arise due to some educators serving vastly different student populations being held to the same expectation. Conversely, if student demographic characteristics are directly controlled for in the prediction models, the concern might arise among stakeholders that some characteristics, such as student race or socioeconomic status, are being used inappropriately for prediction of whether cheating has occurred. The bottom line here is that both possibilities are likely to elicit some form of controversy. The models illustrated in this chapter have typically excluded student demographic covariates due to the aforementioned concerns; however, under certain circumstances, the benefits of including such variables in the model may outweigh potential risks. Consideration of real-world consequences as well as sound judgment informed by a thorough investigation of the test data should be considered essential prerequisites to model building.

Determining the Flagging Criterion

Setting an appropriate flagging criterion to identify unusual changes in performance is a critical decision that requires careful consideration of both theoretical and practical concerns. One noteworthy convenience of standardized statistics, such as the one proposed in this chapter, is interpretability: any researcher with at least a basic understanding of statistics should be able to readily identify classrooms where observed performance greatly deviates from the model's prediction based on the direction and magnitude of the standardized residual. However, it is also critical that the researcher understand and appreciate that normative statistics such as this have their limitations, as the computation of the standardized residual is entirely dependent on the characteristics of the population, and not some absolute, external criterion. As such, the most extreme observed value within any population may appear huge in terms of a standardized residual, although the practical significance of the deviation may be trivial.

Under the right circumstances, it would be entirely reasonable for a researcher conducting investigations on numerous, separate populations to interpret identical values of standardized residuals quite differently, depending on the statistical characteristics of those populations. For this reason, flagging all classrooms or schools with standardized residuals exceeding some predetermined cutoff without any further investigation into the practical significance of such deviations, is not recommended. In some situations, a shift of one student across performance levels in a given classroom, in a given population may look strikingly similar—in terms of the standardized residual—to 10 students shifting performance levels in another population's classroom. Furthermore, flagging criteria may need to be adjusted, depending on the characteristics of the assessment and how the researcher wishes to apply flagging rules. If a single pass/fail

cut point is employed, then applying a uniform flagging criterion (again, with appropriate follow-up review and consideration) is appropriate. However, for tests with more than two possible performance level outcomes, there may be reason to flag units for unusual counts at several different performance levels. For example, if the top two performance level categories are labeled Proficient and Advanced, it might be reasonable to flag classrooms or schools with larger-than-expected proportions of students in *either* category. When applying flagging rules for classrooms with calculated standardized residuals for both of these performance level categories, it is necessary and appropriate to control for Type I error inflation due to multiple significance tests. In this case, the classroom would be flagged if either standardized residual exceeded the flagging criterion associated with $\alpha/2$.

Of course, practical considerations should also be taken into account when determining flagging rules. Based on the size of the population and the alpha level associated with the chosen flagging criterion, the researcher can get an idea as to how many classrooms, schools, and so forth are likely to be flagged in a given investigation. State and local education agencies have finite time and resources to conduct necessary follow-up investigations; thus, they likely cannot investigate every outlying outcome but would most likely focus on those classrooms and schools with the most suspicious outcomes. Flagging rules can be adjusted with this pursuit in mind, or to further simplify the process, standardized residuals could be rank-ordered to prioritize units that appear to be most unusual. Finally, basic common sense should be applied in determining which units are eligible to receive a flag. Units with extremely small student counts, for example, should not be flagged because small deviations from expectation are more likely to result in large standardized residuals.

EMPIRICAL DEMONSTRATION

Description of Test Data

The statistical methodology described in this chapter will now be demonstrated using educational achievement data. In this scenario, longitudinal data in the form of scores and performance levels on a standardized mathematics test from two adjacent years and grade levels were used to form the basis of a practical demonstration of this method. The test was administered in Grade 4 in Year 1 and Grade 5 in Year 2. All students with valid, nonmissing test scores in Year 1 and valid, nonmissing performance levels in Year 2 were included in the model. Missing data is known to be a source of bias in statistical models (Schafer & Graham, 2002). The current example scenario deals with missing scores/performance levels by simply applying listwise deletion, which is acceptable for practical demonstration purposes but is likely problematic for real-world applications, especially if some missing test scores may systematically relate to student achievement and test misconduct. Consideration of test score missingness and potential mitigation strategies will be discussed further in a subsequent section.

Example SAS Code

The following SAS code provides a demonstration of the essential steps required to identify classrooms with unusual student counts at desirable performance levels (e.g., Proficient, Advanced). These essential steps include (1) computing observed student counts at each classroom and performance level for the focal test, (2) using

student-level data in a cumulative logit regression model to predict the focal test's performance level from a prior score and saving individual probabilities output from this model, and (3) computing expected counts, standardizing the difference between observed and expected counts, and flagging classrooms with much greater-than-expected student counts for desirable performance levels. The reader should note the following limitations with the provided code. First, this code is highly customized to a particular data set and scenario. Variable names and some assigned values are tailored to fit this scenario, but other scenarios may require modifications to the code. For example, adapting this approach for an assessment program with three performance level categories at or above Proficient will require the denominator of the critical value adjustment to be modified from 2 to 3, accordingly. Second, this example illustrates the basic concepts of the statistical methodology by predicting students' performance level in the current year from one prior test score; however, using a single prior test score as the sole predictor in such a model is most likely inadvisable for a variety of reasons, which will be dealt with in greater detail subsequently in this chapter.

```
%let CV = 3.00; * Critical flagging value (prior to adjustments).;
%let MinClassN = 5; * Minimum classroom n-count for reporting.;

* Save observed year 2 performance level counts for each
classroom.;

proc freq data = test_data noprint;
  table Y2_ClassID * Y2_PL / sparse
                           outpct
                           out = obs_counts_tall;
run;

* Transpose stacked observed counts into wide format.;

* Compute observed proportions from counts.;

data obs_counts_wide;
  set obs_counts_tall;
  by Y2_ClassID;
  array ObCt [4] ObsCount_PL1 - ObsCount_PL4;
  array ObPr [4] ObsProp_PL1 - ObsProp_PL4;
  retain ObsCount: ObsProp:;
  ObCt[Y2_PL] = COUNT;
  ObPr[Y2_PL] = PCT_ROW / 100;
  if last.Y2_ClassID then output;
  keep Y2_ClassID ObsCount: ObsProp:;
run;

* Fit model and save individual probabilities.;

proc logistic data = test_data;
  model Y2_PL = Y1_MathSS;
  output out = model_probs predprobs = ind;
run;
```

```

* Sum student individual probabilities to get expected counts. ;
proc summary data = model_probs nway;
  class Y2_ClassID;
  var IP_1 - IP_4;
  output out = exp_counts (rename = (_FREQ_ = ClassN)
                            drop = _TYPE_)
    sum = ExpCount_PL1 - ExpCount_PL4;
run;

* Flag classrooms. ;
data final_results;
  merge obs_counts_wide exp_counts;
  by Y2_ClassID;
  array ObPr [4] ObsProp_PL1 - ObsProp_PL4;
  array ExCt [4] ExpCount_PL1 - ExpCount_PL4;
  array ExPr [4] ExpProp_PL1 - ExpProp_PL4;
  array SE [4] SE_PL1 - SE_PL4;
  array SR [4] StdResid_PL1 - StdResid_PL4;
  array Flg [4] Flag_PL1 - Flag_PL4;
  * Adjust flagging for multiple comparisons. ;
  * In this example, the top two PLs will be evaluated. ;
  * Adjust the denominator for more or fewer comparisons. ;
  FlagCritVal = probit(1 - (1 - cdf("NORMAL", &CV)) / 2);
  do i = 1 to 4;
    ExPr[i] = ExCt[i] / ClassN;
    SE[i] = sqrt(ExPr[i] * (1 - ExPr[i]) / ClassN);
    SR[i] = (ObPr[i] - ExPr[i]) / SE[i];
    if i <= 2 then do;
      if SR[i] < -FlagCritVal then Flg[i] = 1;
      else Flg[i] = 0;
    end;
    else if i >= 3 then do;
      if SR[i] > FlagCritVal then Flg[i] = 1;
      else Flg[i] = 0;
    end;
  end;
  FlagSum = sum(of Flag_PL3, Flag_PL4);
  if FlagSum > 0 then Flag = 1;
  else Flag = 0;
  drop i FlagSum;
run;

title1 "Review All Flagged Classrooms";
title2 "Unadjusted Critical Value = &CV / Min N-Count =
&MinClassN";
proc print data = final_results noobs;
  var Y2_ClassID ClassN ObsCount: ExpCount: StdResid:;
  format ExpCount: 5.1 StdResid: 7.2;
  where Flag = 1 and
        ClassN >= &MinClassN;
run;

```

Select Results and Discussion

The methodology described in this chapter was applied to a data set containing student test scores and performance levels across two adjacent years. Select output obtained directly from this analysis is provided. The reader should note that results selected to be discussed in further detail in this section were selected simply to illustrate examples of various potential scenarios and should not be interpreted as being indicative of cheating for any or all of these selected classrooms.

Using the cumulative logit regression model, students' ordinal performance level categories on this year's assessment were predicted from these students' scale scores from the prior year. In this scenario, the implicit expectation is for all students to follow similar growth trajectories, so only fixed effects are included in the model, and no additional adjustments are made to alter student growth trajectories. Student counts by performance level are provided for three classrooms in Table 13.1. The Grade 5 Mathematics assessment has four performance level categories. Performance Level 3 is labeled Proficient and Performance Level 4 is labeled Advanced. Students scoring into either of these categories have satisfied or exceeded grade-level proficiency standards, whereas students scoring into Performance Levels 1 or 2 have not yet demonstrated on-grade proficiency in Mathematics. For flagging purposes, much larger than predicted counts of students at either Performance Level 3 or 4 are flagged. As shown in the provided SAS code, the chosen flagging critical value is set to 3.00 (i.e., classrooms with observed counts at Performance Level 3 or 4 exceeding +3 standard deviations above the expected counts), and this value is further adjusted for multiple comparisons to be approximately 3.205. As shown in Table 13.1, performance level counts for classroom 11 were found to be reasonably consistent with expectations. Classroom 723 was flagged for having more students scoring into Performance Level 3 than would be considered to be reasonable, according to the flagging criteria. Based on testing history, this classroom comprised of a total of 9 students was expected to have 19% of students (or slightly fewer than two) but was observed to have 78% of students (or seven) scoring into this performance level, which resulted in a standardized residual of 4.52. Classroom 1068, which had 20 students, was flagged for having an unusual number of students scoring into Performance Level 4. Based on prior performance, 9% of students (or slightly less than two) were expected to score into Performance Level 4, whereas 30% of students (or six) scored into this performance level, resulting in a standardized residual of 3.26.

Individual results from classroom 723 will be reviewed to illustrate the computation of expected counts and proportions, as well as the standard errors for the four

Table 13.1 Results for Select Units

Unit	N	Observed Proportion				Expected Proportion				Standardized Residual			
		1	2	3	4	1	2	3	4	1	2	3	4
11	21	0.29	0.14	0.57	0.00	0.24	0.18	0.54	0.04	0.44	-0.44	0.29	-0.89
723	9	0.00	0.22	0.78	0.00	0.59	0.22	0.19	0.00	-3.58	0.02	4.52*	-0.19
1068	20	0.05	0.00	0.65	0.30	0.17	0.15	0.59	0.09	-1.44	-1.87	0.55	3.26*

Note: Asterisk indicates that standardized residual associated with a performance level at or above proficient exceed the flagging threshold of 3.00.

Table 13.2 Example Computation of Unit-Level Expected Counts and Proportions

Year 1 Scale Score	Individual Probability			
	1	2	3	4
383	0.97	0.02	0.01	0.00
407	0.77	0.16	0.06	0.00
409	0.74	0.19	0.08	0.00
413	0.66	0.24	0.11	0.00
413	0.66	0.24	0.11	0.00
413	0.66	0.24	0.11	0.00
424	0.40	0.34	0.26	0.00
426	0.35	0.35	0.29	0.00
444	0.09	0.21	0.68	0.02
Expected Count*	5.29	1.98	1.70	0.04
Expected Proportion	0.59	0.22	0.19	0.00
Standard Error	0.16	0.14	0.13	0.02

*Note: Summation occurred prior to rounding for display.

performance levels. Table 13.2 shows the prior year's scale scores for the nine students in this classroom and the estimated individual probabilities for the four performance levels output from the cumulative logit regression model. The expected student count for each performance level is computed by summing individual probabilities at that performance level. Expected proportions are computed by dividing the expected count by the classroom N -count, nine. Finally, the standard error is computed by taking the square root of the expected proportion multiplied by one minus the expected proportion, divided by the classroom N -count.

LIMITATIONS, CAUTIONS, AND FUTURE DIRECTIONS

The modeling strategy presented in this chapter provides a basic framework upon which a researcher can build a functioning longitudinal model to identify classrooms, schools, or districts with unusual changes in performance level classifications relative to prior student achievement. Although the examples included in this chapter are rather simple—with the current year's performance level being predicted by the prior year's test score—this modeling framework can be extended to more complicated modeling scenarios. As previously noted, in a real-world application of such a technique, including a single predictor is inadvisable. Unfortunately, statistical benefits associated with adding more predictors are countered somewhat by burdens associated with more widespread score missingness when additional predictors are included. Simply excluding students who do not have scores for all predictors is inadvisable, and simple approaches such as imputing means for missing scores should also be avoided. Rather, the authors recommend that the researcher should estimate a composite entering achievement score (SAS Institute, 2015; Wright, White, Sanders, & Rivers, 2010) or employ a more sophisticated strategy to deal with missing scores, such as multiple imputation.

The longitudinal modeling method proposed in this chapter is most appropriate for a test that is administered in several adjacent grade levels with at least one prior year of available test data—for example, a standardized end-of-grade assessment for

Mathematics or Reading for Grades 4–8 (assuming the assessment is first administered in Grade 3). For some assessments, such as Grade 3 Mathematics and Reading; end-of-grade standardized tests for other subjects such as Science and Social Studies that are administered only at certain grade levels; and high school standardized end-of-course assessments, finding appropriate predictors may prove to be much more difficult. As is the case with similar (but fundamentally different) value-added models, there is no explicit requirement for all predictors to come from the same subject area as the focal assessment, so long as scores for all predictors are strongly correlated performance on the outcome. For example, it is not necessarily problematic to include prior scores in both Mathematics and Reading in predicting performance levels on the current year's Mathematics test provided that the aforementioned requirements have been met. However, in some circumstances, simply finding a sufficient number of predictors that correlate with an outcome measure of interest may remain problematic. For example, assume that a standardized Science assessment is given to students for the first time in Grade 5. Assuming standardized testing in Mathematics and Reading begins at Grade 3, students may have up to four previous scores to use as predictors of Grade 5 Science achievement. Furthermore, scatterplots and correlation coefficients may reveal strong linear relationships between prior Mathematics and Reading scores and scores on the current year's Science test. However, predicting outcomes on a Grade 5 Science test exclusively from other prior performance in other subject areas such as Mathematics and Reading, may not be desirable. In such circumstances, the researcher must determine if it is most appropriate to proceed using these predictors, to switch to a between-cohort longitudinal modeling design, or to forego longitudinal modeling all together in favor of other statistical techniques to identify unusual patterns in test data.

CONCLUSION

The modeling technique presented in this chapter represents one possible method to be considered when evaluating longitudinal student test data for potentially unusual changes in performance. This technique allows researchers to identify classrooms, schools, or districts with unusual student performance classifications, given their students' prior testing history. When used properly, this gives investigators valuable and potentially actionable information to identify areas requiring appropriate follow-up. However, as noted, in some circumstances other methods may be more appropriate. Common sense, supplemented with sound judgment, informed by a strong understanding of the available data, and consideration of the likely real-world implications of various potential avenues will serve the researcher well in making an informed decision.

REFERENCES

- Agresti, A. (1996). *An introduction to categorical data analysis*. Hoboken, NJ: John Wiley & Sons.
- Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010) Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Educational Policy Analysis Archives*, 18. Retrieved from <http://epaa.asu.edu/ojs/article/view/714>
- Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71, 791–799.
- Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Cizek, G. J. (2001). An overview of issues concerning cheating on large-scale tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, 118, 843–877.

- MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7, 19–40.
- Molland, J. (2012, Oct. 19). *Texas guide to raising test scores: “Disappear” your students*. Retrieved from www.care2.com/causes/texas-cheating-scandal-disappearing-students-to-improve-test-score.html
- National Center for Fair and Open Testing (2013, March 27). *Standardized exam cheating in 37 states and DC; New report shows widespread test score corruption*. Retrieved from www.fairtest.org/2013-Cheating-Report-PressRelease
- National Council on Measurement in Education (2012). *Testing and data integrity in the administration of statewide student assessment programs*. Retrieved from [http://ncme.org/default/assets/File/Committee%20Docs/Test%20Score%20Integrity/Test%20Integrity-NCME%20Endorsed%20\(2012%20FINAL\).pdf](http://ncme.org/default/assets/File/Committee%20Docs/Test%20Score%20Integrity/Test%20Integrity-NCME%20Endorsed%20(2012%20FINAL).pdf)
- SAS Institute (2015). *SAS EVAAS for K-12 statistical models*. Retrieved from: www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/sas-evaas-k12-statistical-models-107411.pdf
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Vogell, H. (2011, July 6). Investigation into APS cheating finds unethical behavior across every level. *The Atlanta Journal-Constitution*. Retrieved from www.ajc.com/news/investigation-into-aps-cheating-1001375.html
- Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010). *SAS EVAAS statistical models*. Retrieved from SAS Institute website: www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf

14

DETECTING UNEXPECTED CHANGES IN PASS RATES

A Comparison of Two Statistical Approaches

Matthew Gaertner¹ and Yuanyuan (Malena) McBride

INTRODUCTION AND PERSPECTIVES

When test scores are used to inform high-stakes decisions—about schools, teachers, or individual students—those who take the tests and those who administer the tests may be tempted to fraudulently improve the scores. The urge to cheat is understandable when serious consequences are attached to assessment results. But to understand a cause is not to condone it. Indeed, the broader impacts of cheating coupled with its increased prevalence threaten to undermine the integrity of educational assessment (Cizek, 2003). Increasingly common reports of test fraud are particularly troubling when we consider that the provision of vital educational services (e.g., supplemental instruction or special education) may be triggered in part by test results. In those cases, to tolerate cheating on the part of the test administrator or test taker is to effectively deny support to those students who need it most (Cizek, 1999).

Using statistical analysis to detect test score irregularities is an important component of a broader test security policy intended to support the accurate measurement of examinees' performance and ensure that test results are meaningful and valid. Statistical methods for detecting test fraud are numerous and can be differentiated to some extent by the type of cheating they are designed to detect. Two categories of methods are described in the following sections.

Item-Level Methods

Some detection methods use item-level response data and focus on examinee-initiated cheating. For example, a variety of methods have been proposed to detect unusual similarities between responses across examinees—the type of irregularities often attributed to one student copying another's answers (Allen, 2012; Hanson, Harris, & Brennan, 1987; Wollack, 2003, 2006; Wollack & Cohen, 1998; Wollack, Cohen, & Serlin,

2001). Answer copying may be the most familiar and convenient approach to cheating on a conventional paper-and-pencil test; not surprisingly, a variety of statistics have been devised to detect it. For example, the K-index (Holland, 1996) focuses on similar patterns of incorrect answers across pairs of examinees, whereas g_2 (Frary, Tideman, & Watts, 1977) and ω (Wollack, 1997) focus on the extent to which the frequency of identical answers (both correct and incorrect) exceeds what would be expected by chance.

Other item-level methods target irregularities whose source is less certain, and whose solution is less clear-cut. In these cases, person-fit analyses (e.g., Drasgow, Levine, & McLaughlin, 1987) offer flexible tools to uncover unexpected strings of correct answers. Karabatsos (2003) provides an extensive list of parametric and nonparametric statistics useful for detecting global misfit—response pattern aberrations that could owe to cheating, careless responding, lucky guessing, creative responding, or random responding. Sijtsma and Meijer (2001), on the other hand, propose the person response function (PRF) as a means to detect *local* aberrations from expected score patterns on selected subsets of items. For example, a researcher analyzing local aberrations may be interested in detecting an unexpected string of correct responses to a particularly difficult set of test questions. When local aberrations are present, student-initiated fraud (e.g., a cheat sheet) may be the culprit, but it is also possible that item responses were altered by test administrators.

Score-Level Methods

Statistical methods that focus on irregularities at the level of total test scores rather than the level of individual items often utilize longitudinal data to detect unexpected changes over time. For example, Jacob and Levitt (2002) developed a cheating indicator that assumes students in classrooms where cheating has occurred are likely to make large score gains in the “cheating year” followed by comparatively small gains or even declines once the cheating stops. Similarly, Simon (2012) employed a regression-based local outlier detection (RegLOD) algorithm to identify schools with outlying scores in the current year relative to peer schools.

Test fraud detection methods based on score changes over time range from the simple to the complex, but at this point it is useful to point out a couple distinctions between score- and item-based detection techniques. First, test scores are economical summaries of item responses: that is, all else being equal, less data and fewer computational resources may be required for score-based fraud detection algorithms than for item-based methods. Second, as evidenced by the examples provided previously, score-based techniques tend to focus on aggregate (i.e., classroom- or school-level) changes, and therefore may be more useful for detecting cheating initiated by test administrators or school leaders, not students. Score-based measures aggregated to the classroom or the school may also be useful metrics for external investigators to uncover test fraud. Any educational agency interested in addressing test security problems would be wise (at minimum) to examine the methods local media outlets will employ to detect those problems. Failure to explore these techniques may lead to embarrassing situations in which newspapers uncover test fraud before state or local education officials are aware of it. For example, in 2012, the *Atlanta Journal-Constitution* conducted a simple study, regressing Atlanta students’ scores in one year on their scores the prior year. Residuals

were aggregated to the school level, and schools with exceptionally high values were flagged (*Atlanta Journal-Constitution*, 2012). The ensuing report was not merely embarrassing for Atlanta Public Schools officials; it led to 35 indictments (including the district's superintendent) and 11 racketeering convictions of other administrators, teachers, and other school personnel (Cook, Niesse, & Tagami, 2015).

When widespread cheating is detailed in popular media, the resulting controversy can spur healthy reforms in test security. States and districts might clarify and vigorously enforce broader fraud deterrence policies. They might also begin exploring the variety of statistical tools available for flagging aberrant score patterns. The experiences of states implementing test security procedures are therefore quite instructive for the field of test fraud detection as it continues to evolve. To that end, this chapter focuses on statistical methods developed and tested in a large U.S. state. The methods discussed herein were designed to identify cheating initiated by educators and school administrators, and we classify these techniques under the score-based rather than item-response-based family of fraud detection strategies. Moreover, because teachers and administrators subject to accountability policies are often evaluated according to the number or percentage of students reaching established proficiency levels, the methods we describe focus not on scale scores but rather on pass rates. Although the research literature covering longitudinal analysis of scale scores is abundant, the literature on pass rate analysis is relatively thin. Thus, we intend for this study to serve as a resource for researchers and practitioners interested in detecting cheating when proficiency classifications drive test-based rewards and sanctions.

BACKGROUND

This section provides some context for our analyses—including a brief description of the test-driven accountability systems that might give rise to suspicious year-to-year changes in pass rates. After all, different accountability metrics create incentives to cheat in different ways, so picking the right fraud detection technique requires considering context. First, however, we offer a cautionary note about the appropriate, circumscribed role of statistical analysis in supporting fair and valid test score interpretations.

The Place of Fraud Detection in a Comprehensive Test Security Policy

In any description of statistical test fraud detection measures implemented in a state-wide assessment program, it is important to emphasize that statistical analyses should be situated in a broader test security policy. Comprehensive security policies and procedures are important for two reasons. First, no statistical detection method can be expected to identify every instance of fraud under every circumstance, so steps focusing on prevention and deterrence are essential. Second, the procedures we describe here may flag incidences of legitimate—albeit, exceptional—learning gains rather than cheating. Test security protocols that fail to recognize this concern will generate substantial public outrage before experiencing a swift and absolute demise. The policies more likely to gain widespread acceptance might well be those that recognize the limitations of statistical detection methods and use these methods primarily to identify schools or classrooms that deserve a closer look.

In the large state where our data originated, a comprehensive plan was introduced to help assure parents, students, and the public that test results are meaningful and valid. Provisions included an analysis of scrambled blocks of test questions to detect copying, required use of seating charts, independent test monitors, and serious consequences (e.g., lowering of an accreditation rating) if cheating is confirmed. Among these procedures is an analysis of score changes over time to detect suspicious gains (or losses, if analysis occurs after cheating cessation). This chapter focuses on change-over-time analytic approaches.

Proficiency Classifications as Outcomes

An initial pilot study (not covered in this chapter) looked for unexpected changes in students' scale scores over time, aggregated to the school level. This method focused on test score residuals—the difference, in scale score units, between what a student achieved and what he or she was predicted to achieve. Our follow-up analysis focused not on students' scale scores, but rather on their binary classifications (i.e., passing/failing) under the state's performance standards. These "pass rate analyses" may be particularly sensitive to small unexpected changes in students' scale scores that nonetheless result in substantial changes in school-level proficiency rates. In other words, pass rate analysis may be more suitable for identifying schools in which test security violations have moved large numbers of students from just below a passing standard to just above it.

School accountability frameworks in this state have typically considered—among many other factors—proficiency rates at specified grade levels across student cohorts (e.g., Grade 10 mathematics performance in 2011, Grade 10 mathematics performance in 2012, and so on). In light of these accountability requirements, school-level cheating may focus on specific grades rather than on student cohorts. As such, our pass rate analyses did not rely on student-level longitudinal data. Rather, each method covered in this chapter relied on cross-sectional data (e.g., Grade 10 pass rates in 2011, Grade 10 pass rates in 2012) to detect efforts to game accountability policies that are focused on the same levels of aggregation.

STATISTICAL MODELS

We evaluated two statistical approaches for conducting pass rate analyses: the two-proportion z -score and multilevel logistic regression. The two-proportion z -score approach examines differences in pass rates across two time points for a group of students (e.g., a school) relative to aggregate pass rate differences across those same time points for all students. Multilevel logistic regression (MLR; Wong & Mason, 1985) examines the relationship between the school a student attends and the likelihood that student will attain a passing score, holding other factors constant. The z -score method requires only school-level data, whereas the MLR requires student-level data. Both methods are detailed below, followed by a brief discussion of each model's strengths and weaknesses.

Two-Proportion Z-Score

Under this approach, each school's z -score is a function of (1) the difference between the school's pass rates in Year 1 and Year 2, (2) the difference between population pass

rates in Year 1 and Year 2, and (3) the size of the school (i.e., student population) in both years (Daniel, 2008). The z -score is calculated as follows:

$$z = \frac{(p_{t2j} - p_{t1j}) - (\bar{p}_{t2} - \bar{p}_{t1})}{\sqrt{\frac{\bar{p}_{t2}(1 - \bar{p}_{t2})}{n_{t2j}} + \frac{\bar{p}_{t1}(1 - \bar{p}_{t1})}{n_{t1j}}}}. \quad (1)$$

In Equation 1, schools are indexed by j ($j = 1, \dots, J$). The pass rate for school j in Year 1 is represented by p_{t1j} , and the pass rate for school j in Year 2 is represented by p_{t2j} . In addition, \bar{p} represents the aggregate population pass rate ignoring school membership (\bar{p}_{t1} for Year 1 and \bar{p}_{t2} for Year 2). Finally, the size of school j in Year 1 is represented by n_{t1j} , and Year 2 school size is represented by n_{t2j} . Large positive z -scores indicate potential cheating in Year 2 (i.e., larger-than-average increases in pass rates between Year 1 and Year 2), whereas large negative z -scores could indicate the cessation of cheating in Year 2 (i.e., larger-than-average decreases in pass rates after test fraud comes to an end).

In subsequent sections of this chapter, we provide an overview of a proposed process for determining which z -scores are large enough to merit further investigation or monitoring. It is worth noting, however, that a measure of practical significance is also available under the z -score approach. The effect size of the difference in pass rates across years is expressed as Cohen's h (Cohen, 1988).

That equation is given below:

$$h = 2 \times \arcsin \sqrt{p_{t2j} - p_{t1j}} - 2 \times \arcsin \sqrt{\bar{p}_{t2} - \bar{p}_{t1}}. \quad (2)$$

Typically, Cohen's h effect sizes between 0.5 and 0.8 are considered moderate, while h values above 0.8 are considered large. Still, effect sizes are difficult to evaluate in isolation, absent consideration of the goals and potential costs of any test security policy. As such, states may not want to rely solely on effect sizes to make determinations about which schools or classrooms need to be investigated further.

Multilevel Logistic Regression

Under this approach, a student's likelihood of attaining a passing score in Year 2 is modeled as a function of (1) the school he or she attends and (2) the prior-year pass rate at that school. The two-level MLR model is specified as follows:

Level 1 (STUDENT):

$$p(y_{ij} = 1) = \frac{e^{\beta_{0j} + r_{ij}}}{1 + e^{\beta_{0j} + r_{ij}}}$$

Level 2 (SCHOOL):

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \times \ln \left(\frac{p_{t1j}}{1 - p_{t1j}} \right) + u_{0j}$$

where $r_{ij} \sim N(0, \sigma^2_r)$, and

$$u_{0j} \sim N(0, \sigma^2_u) \quad (3)$$

In Equation 3, students are indexed by i ($I = 1, \dots, I$) and schools are indexed by j ($J = 1, \dots, J$). At Level 1, the binary variable y_{ij} takes a value of “1” if student i at school j in Year 2 attains a passing score, and “0” otherwise. r_{ij} is a residual term, specific to student i within school j (i.e., the deviation of student i from the school j mean). At Level 2, the Level 1 parameter β_{0j} is expressed as a linear combination of the intercept term γ_{00} , the slope parameter γ_{01} associated with the Year 1 pass rate² at school j (i.e., p_{t1j}), and the residual term u_{0j} . The residual u_{0j} is specific to school j , and it summarizes the association between attending school j and attaining a passing score in Year 2, holding constant school j ’s prior-year pass rate. That residual is the focus of this multilevel analysis; like the z -score detailed above, u_{0j} functions as a school-level measure of performance beyond expectation. In general, large positive values indicate potential test fraud (i.e., larger-than-expected increases in pass rates between Year 1 and Year 2).

A few strengths and weaknesses of each method should be briefly noted here. The most important weakness is emphasized first. Both the z -score and multilevel logistic regression approaches essentially identify schools with better-than-expected Year 2 pass rates. This is not direct evidence of test security violations. It may, in fact, indicate unusually great achievement gains, attributable to particularly effective instructional practice, greatly enhanced student effort, or a combination of many such factors. Consequently, a previous point bears repeating: Sound test security policy is comprehensive in nature, relying on multiple inputs, and informed but not driven by a single statistical index.

Relative to the MLR model, the z -score approach has some attractive features. It is transparent and computationally straightforward. It requires no specialized software for estimation. It is also symmetric: If we seek to identify cheating in Year 2, schools with positive z -scores may be flagged for review, and if we seek to identify cessation of cheating in Year 2, schools with negative z -scores may be flagged for review. In addition, effect sizes can be calculated to summarize practical significance. Last, the z -score formula accounts for school size, thus avoiding the overidentification of small schools common in unadjusted linear regression³ approaches.

Like the z -score approach, the MLR approach implicitly accounts for school size by utilizing a Bayesian shrinkage estimator, whereby relatively low-reliability estimates (sometimes due to few observations) approach zero (Raudenbush & Bryk, 2001). Moreover, unlike the z -score approach, multilevel models can accommodate additional student- or school-level covariates, such as demographic variables.⁴ Multilevel logistic regression, however, is not symmetric. Independent and dependent variables must be swapped if we seek to identify cessation of cheating in prior years.

Research Question

Even though the z -score approach has simplicity and symmetry in its favor, the primary criterion for evaluating statistical methods to support test security should be *accuracy*—the likelihood that a statistical approach will identify a school where test security violations have truly occurred. The more accurate statistical method for this purpose is the approach that detects schools where violations have occurred while minimizing the number of schools that need to be flagged.

For example, imagine a set of 100 schools, one of which is known to have started cheating in Year 2. The known cheater should—if these fraud detection methods are of any use—have higher than average scores under the MLR and z -score approaches. The 100 schools are sorted, descending, once by the MLR approach and again by the z -score

Table 14.1 Descriptive Statistics for Statewide Data and the Samples

	Number of Schools	Pass Rate	Mean Scale Score	Standard Deviation
2011 Statewide	1,824	72.9%	2,189	206
2011 Sample	800	73.8%	2,203	181
2012 Statewide	1,839	72.1%	2,191	212
2012 Sample	800	73.5%	2,206	185

approach. The true cheater is fifth from the top in the MLR-based rankings, and 10th from the top in the z-score-based rankings. The MLR approach wins. In other words, the preferred method should have high power (i.e., a high percentage of true violations are detected), while minimizing Type I errors. This focus on accuracy drove our central, organizing research question: Which technique successfully detects schools where cheating has taken place while limiting the number of schools identified for further monitoring?

Of course, to conduct an accuracy analysis, the analyst needs at least one school where cheating is known to have occurred. To solve this problem, we simulated cheating under a variety of conditions. In the next sections, we describe the available data and the simulation methods.

DATA

This study used 2011 and 2012 Grade 10 mathematics scores from a large statewide assessment program. To accommodate computational resources, a smaller sample of 800 schools was randomly selected from the statewide data. In Table 14.1 we present the number of schools, Grade 10 mathematics pass rates by year, mean scale scores, and standard deviations from both the statewide data and our samples.

SIMULATION METHODS

Preliminary Analysis

To explore the relationships between school characteristics and the relative accuracy of each statistical method, we conducted a preliminary study. The results merit brief discussion here, because our findings motivated a revised simulation design. In this preliminary study, we examined the efficiency of the z-score and MLR methods in detecting various levels of simulated cheating at large and small schools. For the sake of simplicity (and because we doubted it would matter), the simulated cheating schools' Year 1 (baseline) pass rates were set to the population pass rate—roughly 73%.

Two findings surprised us. First, the comparative performance of the statistical methods seemed to depend on the size of the school where cheating was simulated. The z-score approach was more accurate when cheating occurred at a large school, whereas the MLR method was more effective detecting cheating at small schools. Second, both methods proved startlingly inaccurate. For example, to flag any school (large or small) where 30% of failing scores had been fraudulently changed to passing, 50% of the schools in the state would have to be flagged. Even if we were willing to look for cheating only at large schools (where the z-score method is most accurate), 13% of the schools in the state would need to be flagged.

Our results suggested that (1) neither the *z*-score nor the MLR methods held much promise for test fraud detection, given the resources available for monitoring flagged schools; and (2) it might be useful to establish a “dividing line” between large schools and small schools, to determine the precise point along the school-size continuum where the *z*-score method becomes more accurate than MLR approach. Therefore, in the final simulation design, school size was the primary focus. We also varied cheating schools’ baseline pass rates relative to the population pass rate, but we took this step essentially as a robustness check; setting cheating schools’ Year 1 pass rates to the population pass rate in the preliminary study was a somewhat arbitrary choice.

Final Simulation Conditions

Our revised simulation design allowed comparison of the *z*-score and MLR approaches under a variety of conditions, including different-sized cheating schools (from small to big), different Year 1 pass rates at cheating schools and different population pass rates (from low to high), and different amounts of cheating (from minor to blatant). In each condition, one cheating school was simulated—increasing the sample size from 800 schools to 801.⁵ We varied the size of the 10th-grade class in the simulated cheating school, between 50 and 500 in increments of 50. Year 1 (i.e., 2011) pass rates also varied, from 10% to 90% passing in increments of 10 percentage points.

The extent of cheating was also manipulated, and this simulation step deserves a more detailed explanation. Each student in the dataset has a passing indicator, set to “1” if a student passed the 2012 (Year 2) Grade 10 mathematics test and “0” if the student failed. The “true” (fraud-free) performance of the cheating school in any given condition was assumed to be equivalent pass rates in Year 1 and Year 2. In one condition of the simulation, 10% of the zeros (indicating a failing score) in Year 2 at the cheating school were set to “1” (indicating a passing score).⁶ This step was meant to simulate a scenario where, owing to test fraud, 10% of students would be indicated as passing even though their “true performance” was below the passing standard. In the next condition, 20% of failing scores were set to passing scores. In the next condition, that percentage increased to 30%, and so on. Thus, the extent of simulated cheating ranged from 10% to 100%, in increments of 10 percentage points.

Finally, to investigate the impact of differences between school-level and population pass rates, we included four population pass rate conditions—at approximately 30%, 50%, 70%, and 90%. The various permutations of school size, Year 1 pass rate, extent of cheating, and the population pass rate resulted in 10 (school sizes) × 9 (Year 1 pass rates) × 10 (extents of cheating) × 4 (population pass rates) = 3,600 different simulated conditions. Comparing the accuracy of each statistical method under each condition was straightforward. We examined the percentage of high schools sample wide each method needed to flag to catch the simulated cheater.⁷ The method judged to be preferred was defined as the one that detected the cheating school while flagging a smallest percentage of schools in doing so.

RESULTS

All the variables we manipulated—the extent of cheating, the initial-year pass rate and the population pass rate, and school size— influenced the accuracy of the *z*-score and MLR approaches. It was not surprising that the extent of cheating affected the likelihood that either statistical approach would detect it. Generally, and perhaps obviously,

when more blatant levels of cheating were simulated, both the *z*-score and MLR approaches flagged fewer schools before catching the test security violation.

Interestingly, however, school size had a much smaller impact than we had anticipated. The discussion of our preliminary analysis above (focused on school size) was therefore more than an elaborate detour. We meant to illustrate that a seemingly important effect under one simulation design essentially vanished when other factors were taken into account. In sum, the school size effects on accuracy are dwarfed by the interaction of baseline and population pass rates. Regardless of school size, when the Year 1 pass rate of a simulated cheating school is substantially below the population pass rate, the *z*-score method is a much more efficient approach. Conversely, when the baseline pass rate is higher than the population pass rate, MLR is more accurate. In simple terms, catching low-performing cheaters is easier with the *z*-score approach, and catching high-performing cheaters is easier with the MLR method.

This finding, along with the “extent of cheating” effect, is illustrated in Figure 14.1. Here, we show the accuracy of each method as a function of the extent of cheating and school size. Accuracy is defined as the number of schools each method had to flag before catching the cheater; smaller is better. The population pass rate in Figure 14.1 is approximately 50%, and the Year 1 pass rate for the simulated cheater is 10%. For ease of presentation, we present three school sizes—50 students in the 10th-grade class, 250 students, and 500 students.

First, the “extent of cheating” effect is illustrated plainly in Figure 14.1. Under either method, the more often a cheating school changed scores from failing to passing, the easier it was to identify. The *z*-score approach, however, is consistently more efficient, because the Year 1 pass rate of the simulated cheating school (10%) is so far below

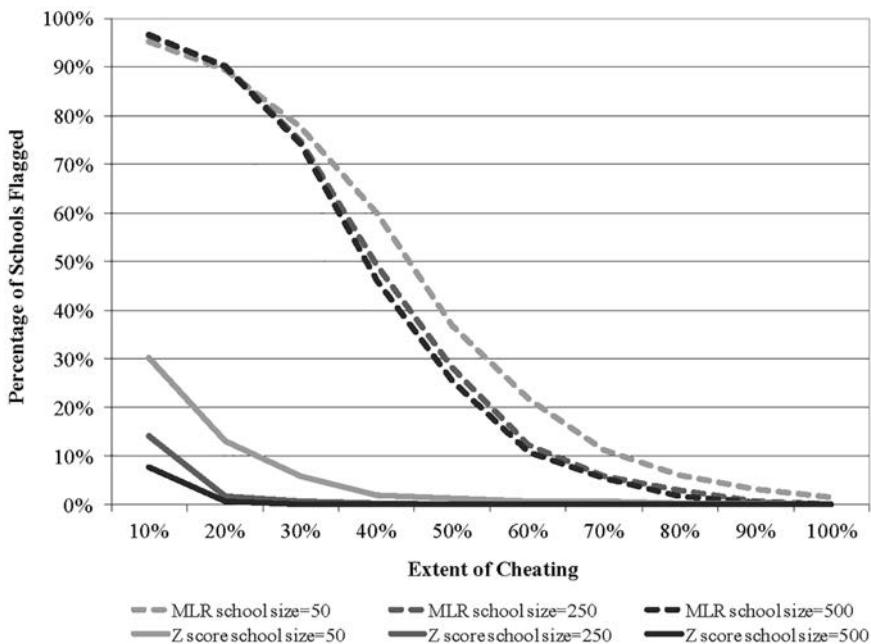


Figure 14.1 Comparative Accuracy of *z*-Score and MLR Approaches, by Extent of Cheating and School Size (baseline pass rate = 10%; population pass rate = 50%)

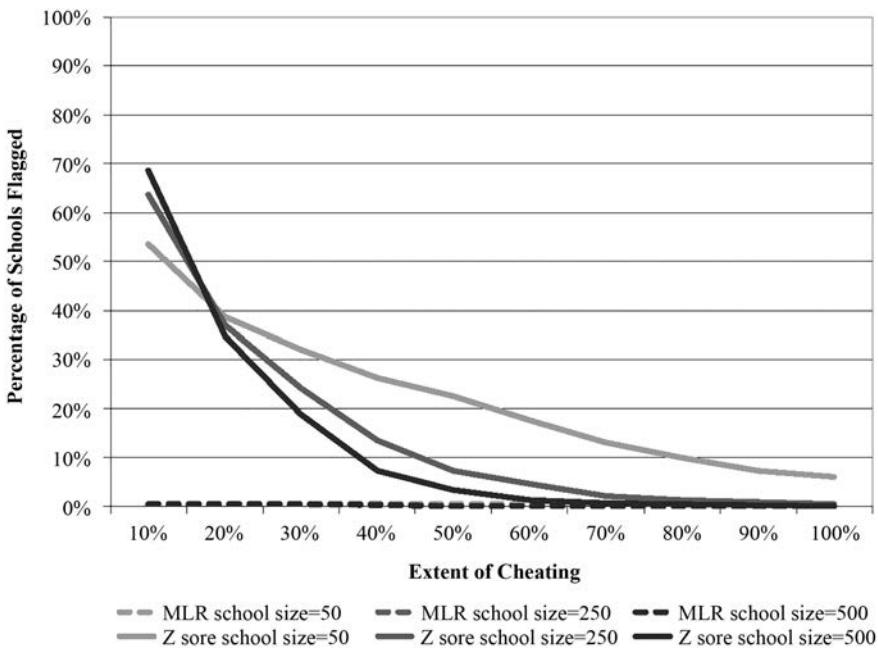


Figure 14.2 Comparative Accuracy of z-Score and MLR Approaches, by Extent of Cheating and School Size (baseline pass rate = 80%; population pass rate = 30%)

the population pass rate (50%). We illustrate the opposite effect in Figure 14.2. This time, imagine a population pass rate at 30%, and a Year 1 pass rate for the simulated cheater at 80%.

Figure 14.2 reinforces the “extent of cheating” effect on accuracy. More blatant cheating leads to more efficient detection. More important, Figure 14.2 depicts how the interaction of population pass rates and Year 1 pass rates for cheating schools can affect judgments about statistical methods’ comparative accuracy. Here, the MLR approach is the clear winner—consistently more efficient than the z-score method in detecting test fraud. In most cases, MLR needs to flag only two or three out of 801 schools before identifying the simulated cheater.

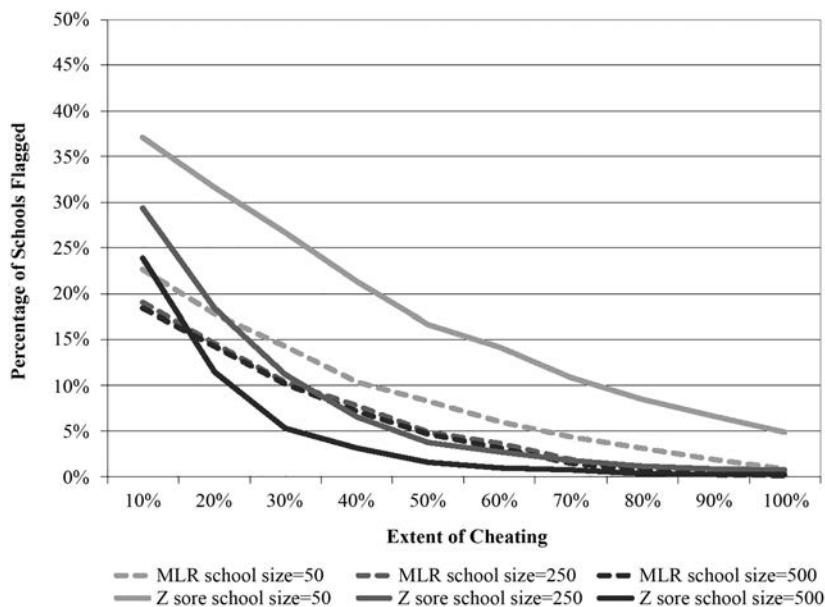
In Table 14.2, we summarize the full complement of simulation results, with one cell for each combination of baseline and population pass rate. Our method of synthesis in Table 14.2 is simple. Each combination of baseline and population pass rate has 100 “z-score versus MLR competitions”—10 extents of cheating (from 10% to 100%) × 10 school sizes (from 50 to 500). In Table 14.2, we highlight the method that wins the highest percentage of those competitions.

Table 14.2 more comprehensively depicts the interaction effect of population and baseline pass rates. If the population pass rate is held constant, increasing the baseline pass rate at a cheating school (i.e., moving from left to right in Table 14.2) increases the comparative accuracy of the MLR method. If the baseline pass rate is held constant, increasing the population pass rate (moving from top to bottom in Table 14.2) increases the comparative accuracy of the z-score approach. Only when baseline and population pass rates are relatively close (the shaded cells in Table 14.2) does this effect diminish. In these cases, school size may be a deciding factor in model accuracy.

Table 14.2 Comparative Accuracy of z-Score and MLR Approaches, by Population Pass Rate and Baseline Pass Rate

		Baseline Pass Rate								
		10%	20%	30%	40%	50%	60%	70%	80%	90%
Population Pass Rate	30%	Z score wins	Z score wins	Z score wins	Z score wins	Z score wins	MLR wins	MLR wins	MLR wins	MLR wins
	83%	81%	74%	57%	38%*	54%	79%	98%	100%	
	50%	Z score wins	Z score wins	Z score wins	Z score wins	Z score wins	MLR wins	MLR wins	MLR wins	MLR wins
	92%	91%	92%	90%	84%	50%	82%	100%	100%	
	70%	Z score wins	Z score wins	Z score wins	Z score wins	Z score wins	Z score wins	Z score wins	MLR wins	MLR wins
	93%	92%	94%	92%	93%	90%	89%	63%	100%	
	90%	Z score wins	Z score wins	Z score wins	Z score wins	Z score wins	Z score wins	Z score wins	MLR wins	MLR wins
	94%	92%	95%	92%	95%	92%	94%	92%	72%	

* Note: Thanks to ties, one method can win less than half the time but still more often than the other method. Cells are shaded where the baseline pass rate is within 10 percentage points of the population pass rate.

**Figure 14.3** Comparative Accuracy of z-Score and MLR Approaches, by Extent of Cheating and School Size (baseline pass rate = 80%; population pass rate = 70%)

School Size

Under both statistical methods and regardless of the relationship between baseline and population pass rate, large cheating schools were typically easier to detect than small ones. Under certain scenarios, school size may also dictate the choice of a test fraud detection method. One such scenario is illustrated in Figure 14.3. Here, the cheating school's Year 1 pass rate (about 80%) and the population pass rate (about 70%) are close, so neither the MLR nor the z-score method distinguishes itself as the clear winner in terms of accuracy.

In Figure 14.3, *z*-score lines (solid) tend to “sandwich” the MLR lines (dashed), at most points along the “extent of cheating” continuum. The sandwich effect may suggest the relevance of school size when baseline and population pass rates are similar: Under these conditions, MLR is generally more accurate detecting small cheating schools, and the *z*-score approach works better for large schools.

PUTTING FRAUD DETECTION METHODS INTO PRACTICE

Our simulation study yielded three general lessons: (1) Low-performing cheaters (those with low Year 1 pass rates relative to the population pass rate) are best detected via the *z*-score approach, whereas high-performing cheaters are best detected via MLR; (2) among average-performing cheaters, large schools are best identified via the *z*-score approach, whereas small schools are best identified via MLR; (3) contrary to preliminary study conclusions, the MLR and *z*-score approaches *do* seem to hold promise as fraud detection tools, under certain circumstances.

Up to this point, our chapter has focused exclusively on the comparative performance of two statistical methods. In the next subsection we will discuss how the models’ outputs may be interpreted and used in practice. In doing so, we hope to illustrate not only the conditions under which the *z*-score and MLR approaches can be powerful test fraud detection tools, but also the cases where abundant caution is warranted.

Establishing Thresholds for Cheating

Under either statistical method, the central question policy makers will face is how much aberrant performance is enough to trigger further investigation. One option would be to standardize the estimates produced by fraud detection models and scrutinize values beyond certain thresholds (e.g., three standard deviations from the mean). Unfortunately, this is not an ideal option. Such thresholds are arbitrary and entirely norm referenced, and interpreting values outside them is difficult if users do not understand the levels of cheating those values represent. Instead, what we propose here is a system designed to indicate the proportion of schools that must be identified in order for various degrees of cheating to be detected.

Figures 14.1–14.3 illustrate a statistical truism: To detect schools engaged in relatively small degrees of cheating, a wider net must be cast. Conversely, to detect only those schools where high degrees of cheating are taking place, the search can be narrowed significantly. Realistic test security policy involves trade-offs. A state agency will not be able to investigate every school in its jurisdiction, so it must weigh available resources against the desire to detect any cases where any test scores have been altered. The necessary exercise therefore involves balancing identification rates (i.e., the percentage of schools that will be flagged for further monitoring) against the level of aberrant performance considered suspicious. Letting the former get too high will strain resources. Letting the latter get too high effectively condones cheating, and may lead to controversy if fraud is detected by investigations initiated by persons other than the relevant education officials. The relationship between the proportion of schools identified and the level of cheating detectable is illustrated in Table 14.3. In this example, the population pass rate is 50%. The hypothetical cheaters’ pass rate is about 70%, so MLR-based detection is preferred.

Using the percentages in Table 14.3, policy makers can make informed decisions, balancing the level of monitoring feasible with the degree of cheating the state would

Table 14.3 Guidelines for Balancing Cheating Thresholds and Identification Rates (MLR)

To detect schools whose baseline pass rate is 70% and where . . .									
10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
. . . of failing scores have been set to passing, you must flag . . .									
7%	6%	5%	3%	2%	2%	1%	1%	<1%	<1%
. . . of all schools in the state.									

Table 14.4 Guidelines for Balancing Cheating Thresholds and Identification Rates (z-Score)

To detect schools whose baseline pass rate is 30% and where . . .									
10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
. . . of failing scores have been set to passing, you must flag . . .									
22%	6%	2%	1%	1%	<1%	<1%	<1%	<1%	<1%
. . . of all schools in the state.									

like to detect. Although it may be desirable to detect instances where 10% of failing scores have been altered, the state may not have the resources to effectively monitor 7% of its schools. A state may instead prefer to monitor 2% of its schools to ensure detection of instances where 50% of failing scores have been altered. This tactic would recognize limited monitoring resources while still ensuring that the worst offenders would be identified. It is worth noting that a level of cheating at 50% is not far-fetched. Using this example, if 50% of failing scores were fraudulently set to passing, test fraud would bump a school's pass rate from 70% to 85%—a percentage-point change that might not, on its face, raise many eyebrows. In Table 14.4, we provide a similar visual representation, this time for cheating schools with a baseline pass rate of 30% relative to a population pass rate of 50%. In this circumstance, the z-score method is preferred.

Again, representations like the one provided in Table 14.4 may help policy makers consider test security in the context of limited resources. It may be difficult to effectively monitor 22% of the schools in a state, in which case detecting cheating levels at 10% would not be feasible. On the other hand, a state could catch the most blatant offenders (cheating levels $\geq 50\%$) while only flagging 1% of schools statewide.

Unfortunately, not every circumstance will yield such efficient detection methods. Imagine a set of schools where baseline pass rates are identical to population pass rates—both at 50%. Many more schools must be flagged to detect the same levels of cheating displayed in Tables 14.3 and 14.4. For example, if a state wanted to catch any school (large or small) that changed 50% of its failing scores to passing, 25% of the schools would need to be flagged. Indeed, when baseline and population pass rates are close, abundant caution is warranted. In such cases a state may wish to use the z-score method and focus only on large schools where cheating may have occurred. Small schools would be given a pass, but to catch cheating at the 50% level, only 3% of schools statewide would need to be flagged.

Handling Two Detection Methods

Advocating the MLR approach in some circumstances and the z-score in others means applying different detection tools to different schools on the basis of prior performance.

Table 14.5 Guidelines for Balancing Cheating Thresholds and Identification Rates (comparing MLR and z-score, population pass rate = 50%)

To detect schools whose baseline pass rate is 70% and where ...										
10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
... of failing scores have been set to passing, under MLR you must flag ...										
7%	6%	5%	3%	2%	2%	1%	1%	<1%	<1%	
... of all schools in the state. But with the z-score you must flag ...										
52%	42%	35%	30%	24%	21%	17%	13%	11%	8%	

Such a policy would need to be crafted delicately. Still, by incorporating a multimethod approach, a great deal of efficiency is gained. Let us return to Table 14.3, where the MLR approach is preferred, and consider an alternative policy where the z-score method is the lone detection tool.

The alternative is not terribly attractive. To detect egregious cheating (say, at the 50% level), nearly a quarter of the schools in the state with the z-score would need to be flagged, compared to 2% of schools statewide under MLR. Of course, when baseline pass rates are far below the population pass rate, the opposite effect is evident: The z-score is clearly the more attractive option. Quite simply, choosing either the z-score or MLR as the sole cheating detection mechanism would force a state to choose the type of school—either high performing or low performing—where it would prefer to detect fraud.

The purpose of presenting Table 14.5 is to suggest that this strategy for establishing detection thresholds—viewing identification rates in concert with cheating thresholds—may stimulate useful conversations about the importance of a multimethod approach to test fraud detection. Simulations and visuals such as Table 14.5 may help states advance the argument that different methods applied to different schools makes the fraud detection process more efficient and, in fact, more equitable; states will not need to choose whether to focus on high-performing or low-performing schools. Moreover, the process we outline in this section may provide a more informative and flexible framework for establishing detection thresholds than would be available with purely normative thresholds (e.g., all z-scores greater than 3).

APPLYING PASS RATE ANALYSIS TO THE COMMON EDUCATIONAL DATASET

Our final analytic step involved using the MLR and z-score approaches to identify possible test security violations in the common educational dataset referenced throughout this book. As with any real-world application of statistical fraud detection methods, flagging potential cheaters in the common educational dataset required a few assumptions and decision rules. First, we needed to put ourselves in the shoes of state officials and decide what level of aberrant performance would be considered suspicious. We assumed that a state education department has relatively limited resources for monitoring flagged schools and would like to focus its fraud detection efforts on particularly suspicious cases. Therefore, we set detection thresholds that would identify schools where at least 50% of failing scores had been set to passing.

Detection thresholds for the MLR and z-score approach were set using our simulation-based guidelines (see Tables 14.3 and 14.4), but these guidelines needed to be

refined to reflect the common educational dataset. Specifically, Tables 14.3 and 14.4 suggest a hypothetical population pass rate of 50%, a “high-performers” pass rate of 70%, and a “low-performers” pass rate of 30%. The common educational dataset looked a bit different. First, the Grade 5 Year 2 population pass rate was 46%, not 50%. Second, the typical high-performing school in the common dataset had a 58% pass rate (that is the median pass rate for all schools with pass rates at or above 46%). Finally, the typical low-performing school in the common dataset had a 31% pass rate (that is the median pass rate for all schools with pass rates below 46%). Given these adjustments, the detection guidelines in Tables 14.3 and 14.4 were refined:

1. To detect schools whose baseline pass rate is 58% and where at least 50% of failing scores have been set to passing, you must flag the top 2% of all schools in the state under the MLR metric.
2. To detect schools whose baseline pass rate is 31% and where at least 50% of failing scores have been set to passing, you must flag the top 0.35% of all schools in the state under the *z*-score metric.

Using these guidelines, 23 schools in the common dataset were identified for further investigation. For ease of interpretation and to help prioritize the list, we also divided this group of 23 schools into two categories. Twenty-one schools were flagged by *either* the MLR or the *z*-score method, but not by both. Another two schools were flagged by *both* metrics. The group of 21 schools flagged by only one metric may be considered medium-priority cases, subject to ongoing performance monitoring and an audit of test security practices. The two schools flagged by both metrics, however, should be considered high-priority cases, subject to intensive monitoring, erasure analysis, interviews with school staff, and independent proctors for all high-stakes test administrations. Why place the highest priority on schools flagged by both metrics? Recall that across our simulation conditions and detection methods, one factor always leads to highly efficient detection: more blatant cheating. Therefore, schools flagged by both metrics are the most likely to have engaged in extensive test fraud, and therefore deserve a closer look.

Whereas these detection guidelines and classification heuristics may be helpful for states with limited monitoring resources, it is fair to point out that “at least 50% of failing scores set to passing” may be viewed as a fairly permissive standard. A state may instead wish to detect schools where, for example, 30% of failing scores have been set to passing. Of course, that choice involves a trade-off: More schools will be identified. In the common educational dataset, the number of schools identified more than doubled. With the cheating threshold lowered to 30%, 66 schools (rather than 23) were identified for further monitoring. Sixty-one schools were flagged under either the MLR or the *z*-score metric (but not both), and five schools were flagged by both metrics.

LIMITATIONS AND FUTURE RESEARCH

The limitations to this study, and the further research those limitations suggest, fit into two categories: methodological and policy related. We discuss each in turn below.

From a methodological standpoint, we may have shortchanged the potential of the MLR approach, in that it has some features that were not fully utilized in our simulation. For example, to avoid privileging either method, both the *z*-score and MLR

approach in this study relied on equivalent inputs (binary passing indicators for Grade 10 mathematics students in 2011 and 2012). Whereas the z-score method is necessarily limited to analysis of unconditional pass rates, multilevel models can accommodate school- and student-level covariates to improve predictive accuracy. The addition of such covariates could improve the efficiency of the MLR approach. In other words, given equivalent data, the preferred method seems to depend on a cheating school's baseline pass rate; given additional covariates (upon which only the MLR could capitalize), the relative strengths of the multilevel approach may be more apparent.

From a policy standpoint, there may be a single-measure alternative that could alleviate concerns about applying different detection methods to different schools as a function of prior performance. More specifically, a state might calculate a z-score and MLR estimate for each school and combine the two measures in a weighted composite. The relative weight for each measure (the z-score and the MLR residual) would depend on the difference between that school's prior-year pass rate and the population pass rate. Large negative values would produce high weights on the z-score measure; large positive values would produce high weights on the MLR measure. Values close to zero would produce roughly equivalent weights for each measure, at which point a state may wish to incorporate a small school size adjustment. Our study did not incorporate such a composite, but future analyses might consider it. Although this may seem like a methodological issue, we consider it policy-relevant because a single composite may help ease a state's concerns about applying different statistical standards to different schools.

Finally, and most important, any discussion of statistical measures to support test security policy should be couched in need for multiple measures and safeguards against unwarranted sanctions. The z-score and MLR methods detect extraordinary growth in pass rates, which does not suffice as proof of test security violations. No behavior (teacher, administrator, or otherwise) is observed. As such, each of the percentages on the bottom rows of Tables 14.3 and 14.4 will inevitably include false positives—schools where no violations have occurred but where statistical models have nonetheless identified anomalous performance from one year to the next. Improvements in student performance may actually owe to improvements in instruction, increased student effort, or other factors, so cheating detection methods based on changes in categorical classifications (e.g., pass rates) should be interpreted with caution. Their use should take place within a larger investigative process that includes the collection of additional evidence, such as locally maintained seating charts, reports of testing irregularities, and test participation data. A single statistical snapshot may trigger a closer look, but it should not be used in isolation to combat test fraud.

This said, we should acknowledge “additional evidence” may be scarce. There may be no witnesses to the fraud nor explicit documentation of it. Under such circumstances, statistical evidence could be strengthened via multiple measurements. Imagine a school where students seem to make extraordinary proficiency gains in a single year. That school is flagged by (for example) the MLR metric, and the following year, independent test proctors are assigned to monitor the assessment administration and ensure test booklets are secure until they arrive at a scoring facility. In that subsequent year, pass rates decline precipitously and the school is again flagged, this time by large negative values on the MLR metric. That pattern—a flag for cheating, followed by strict monitoring, followed by a flag for cessation of cheating—may provide compelling statistical evidence for either outrageous coincidence or, more likely, test fraud.

NOTES

1. This research was completed when Matthew Gaertner worked in the Center for College and Career Success at Pearson.
2. To satisfy linearity assumptions, Year 1 school-level pass rates were converted to the logit scale by taking the natural log of the unconditional odds of attaining a passing score at each school.
3. In addition to the *z*-score and MLR, we estimated a simple linear regression model. This method proved overly sensitive to school size—small schools were disproportionately likely to be flagged for review. As such, the linear regression approach was not pursued further.
4. The MLR model could accommodate more independent variables, such as student demographics or prior-year scale scores. However, those data are not available for the *z*-score method, so student- and school-level covariates were discarded to avoid privileging the multilevel approach with additional data.
5. It is important to keep in mind that the simulated cheater was added to 800 real schools. It is impossible to know whether actual tampering took place at any of those 800 schools. Therefore, in this simulation, we are seeking the fraud detection method that can most efficiently identify a *known* cheater.
6. Because neither statistical method in this simulation considered students' raw or scale scores, 10% of 0s were selected at random to be set to "1" (passing).
7. Although the 800 randomly sampled schools may include a handful where true tampering occurred (i.e., the simulated cheater might not be the only cheater in the data), no systematic bias is introduced favoring the MLR or *z*-score approach, because both methods are tested on the same set of 801 schools.

REFERENCES

- Allen, J. (2012). *Relationships of examinee pair characteristics and item response similarity*. Paper presented at the Conference on Statistical Detection of Potential Test Fraud, Lawrence, KS.
- Atlanta Journal-Constitution (2012). *Cheating our children: The AJC's methodology behind suspicious school test scores*. Retrieved from the Atlanta Journal-Constitution: www.ajc.com/news/news/local/cheating-our-children-the-ajcs-methodology-behind-/nQSTN/
- Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Erlbaum.
- Cizek, G. J. (2003). *Detecting and preventing classroom cheating: Promoting integrity in assessment*. Thousand Oaks, CA: Sage.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cook, R., Niesse, M., & Tagami, T. (2015, April 14). Judge gives convicted Atlanta school cheaters a lesson. *The Atlanta Journal-Constitution*. Retrieved July 1, 2015 from www.ajc.com/news/news/local-education/judge-gives-convicted-atlanta-school-cheaters-a-le/nktSL/
- Daniel, W. (2008). *Biostatistics: A foundation for analysis in the health sciences* (9th ed.). New York, NY: John Wiley and Sons.
- Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, 11, 59–79.
- Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, 6(2), 152–165.
- Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying* (Research Report Series No. 87–15). Iowa City, IA: American College Testing Program.
- Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: statistical theory and empirical support* (ETS Technical Report No. 96-4). Princeton, NJ: Educational Testing Service.
- Jacob, B. A. & Levitt, S. D. (2002). *Catching cheating teachers: The results of an unusual experiment in implementing theory*. National Bureau of Economic Research Working Paper 9414. Retrieved from NBER: www.nber.org/papers/w9414.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16, 277–298.
- Raudenbush, S., & Bryk, A. (2001). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.
- Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, 66, 191–208.
- Simon, M. (2012). *Local outlier detection in data forensics: Data mining approach to flag unusual schools*. Paper presented at the Conference on Statistical Detection of Potential Test Fraud, Lawrence, KS.

- Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement, 21*, 307–320.
- Wollack, J. A. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement, 40*, 189–205.
- Wollack, J. A. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education, 19*, 265–288.
- Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22*, 144–152.
- Wollack, J. A., Cohen, A. S., & Serlin, R. C. (2001). Defining error rates and power for detecting answer copying. *Applied Psychological Measurement, 25*, 385–404.
- Wong, G. Y., & Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association, 80*, 512–523.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Section III

Theory, Practice, and the Future of Quantitative Detection Methods



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

15

SECURITY VULNERABILITIES FACING NEXT GENERATION ACCOUNTABILITY TESTING

*Joseph A. Martineau, Daniel Jurich,
Jeffrey B. Hauger, and Kristen Huff*

INTRODUCTION

Test security has been a scholarly concern for many decades (Cizek, 1999). Cheating behavior and other security breaches can call into question the validity of decisions made on the basis of examinees' scores. To protect the validity of those scores and the credibility of resulting decisions, it is important not only to deter such behaviors but also to detect and respond to them. This chapter considers test security in the context of recent developments in accountability testing, particularly focused on assessments based on new and more rigorous content standards focusing on higher level skills, such as constructing an argument and critiquing and analysis. Recent examples include the Common Core State Standards (National Governors Association & Council of Chief State School Officers, 2010) or similar standards such as those of Texas and Alaska (Conley, Drummond, de Gonzalez, Seburn, Stout, & Rooseboom, 2011; Haymes, 2013), and the Next Generation Science Standards (Achieve, Inc., 2015).

One key implication of more rigorous content standards is that tests must include more items measuring higher level skills than have been common in state assessment programs for the past decade, as witnessed by proposals for new assessment systems based on more rigorous standards (PARCC, 2010; Smarter Balanced, 2010). Both proposals include a large number of constructed response items and performance tasks that go beyond the more conventional multiple-choice response formats that have dominated state testing programs.

Although the Smarter Balanced and PARCC assessments move many paper-based state assessments online, large-scale computer-based testing (CBTs) with high stakes for test takers has been around for quite some time. The Graduate Record Examination was first field-tested as a CBT in 1991 (Schaeffer, Reese, Steffen, McKinley, & Mills, 1993), and was used operationally as a computer-delivered test for the first time in 1993 (Mills & Steffen, 2002). On the one hand, large-scale CBTs with high stakes for educational entities and individual educators is a relatively new phenomenon within

the last decade (Poggio & McJunkin, 2012), although CBT was already being implemented in some fashion in over half the states by 2011 (Martineau & Dean, 2012). On the other hand, for many states, the 2014–2015 school year will be the first operational, large-scale use of online assessment in a high-stakes context. These developments in next generation testing pose considerable new challenges in test security to add to the old challenges. Some reasons for these challenges are described below.

ACCOUNTABILITY AND HUMAN BEHAVIOR

As predicted by educational researchers at its inception (Goertz & Duffy, 2003), the enactment of NCLB (No Child Left Behind Act, 2001) ushered in an era of accountability testing that increased both the amount of testing and the stakes associated with tests in the U.S. educational system. In this accountability-focused environment, the federal government, State Education Agencies (SEAs), and Local Education Agencies (LEAs) utilize results from educational assessments for various high-stakes and potentially life-altering purposes. For example, students' progression into the next grade (e.g., Florida Department of Education, n.d.; Ohio Department of Education, n.d.) or to graduation (e.g., Georgia Department of Education, n.d.; New York State Education Department, 2010) can depend on passing a single or several exams. In many states, teacher and school administrator job performance evaluations now include student test performance as a component of the overall evaluation (Hull, 2013). NCLB and associated programs have tied federal programmatic as well as monetary outcomes to school and district performance on state assessments.

Before the implementation of NCLB, Chicago Public Schools (CPS) used test results for high-stakes purposes. Based on their analysis of CPS test data, Jacob and Levitt (2003) concluded that high-stakes use of test results had resulted in cheating behavior by teachers/administrators in a minimum of 4% of CPS classrooms. This is a high enough percentage to call into question aggregate CPS results, let alone results for individual schools, classrooms, and students. Although a causal link is unproven, the implementation of NCLB and associated programs has been followed by some large-scale test security breaches. Few have garnered national attention as strongly as recent scandals involving adult cheating on standardized tests in Atlanta, Georgia (Perry & Vogell, 2008) and El Paso, Texas (Martinez & Anderson, 2016). Reports published in the *Atlanta Journal Constitution* were followed up by investigative reporting across the nation, which concluded the following:

- statistical analyses of large gains or drops provide evidence of widespread systematic cheating across many school systems in many states (Perry, Vogell, Judd, & Pell, 2012);
- states have inadequate procedures in place to discourage and investigate cheating, which results in cheating continuing to thrive (Judd, 2012); and
- more cheating scandals are inevitable because states are incapable of ensuring test integrity (Pell, 2012).

These articles on the Atlanta cheating case resulted in a flurry of follow-up articles by national media outlets, bringing the story to a wider audience, including a publication (Toppo, Amos, Gillum, & Upton, 2011) flagging potential cheating by educators in all states where the authors conducted analyses (Arizona, California, Colorado, Florida, Michigan, Ohio, and Washington, DC). National Public Radio later picked up the El Paso story (Sanchez, 2013), bringing it also to national attention, and serious

concerns about large-scale cheating on standardized tests were reported for Washington, DC (Toppo, 2013), and Philadelphia (Graham, 2014).

In response to the Atlanta cheating scandal in particular, the United States Secretary of Education contributed an article to the *Washington Post* (Duncan, 2011) acknowledging the problem of cheating, but defending the need to continue to measure student achievement and student growth and to hold educators accountable through those means. He also referenced the grants provided by the U.S. Department of Education (USED) to consortia of states to develop next generation assessments that would improve the quality of the assessments (2010b). Also in response to the cheating scandals and to a request from USED, the National Council on Measurement in Education published a document on test integrity (2012), providing guidelines on improving and monitoring test integrity.

The need for improvements in test security is likely to continue. Recent developments in education policy have further increased the stakes attached to test results. For example, implementation of school improvement grants (U.S. Department of Education, 2010a) under the American Recovery and Reinvestment Act of 2009 and the implementation of the ESEA flexibility waiver program (U.S. Department of Education, 2013a), principals leading the worst performing schools lose their jobs except under limited circumstances. In addition, as debate began on reauthorizing NCLB (or the Elementary and Secondary Education Act of 1965, or ESEA), it became clear that a compromise would not be reached in time to change the NCLB/ESEA requirement that 100% of students be proficient in reading and mathematics by 2014. The U.S. Secretary of Education then acted on his authority under NCLB to grant waivers from certain provisions of NCLB to states to avert the consequences of arriving at that point (U.S. Department of Education, 2013a). USED developed a process by which states could request waivers if certain requirements were met (U.S. Department of Education, 2012). Those requirements include the development and implementation of educator evaluations based in part on student achievement/growth data, where the evaluation results must be used for high-stakes decisions regarding educators such as hiring, firing, retention, promotion, and compensation.

Given the increasingly critical decisions informed by educational testing results, it is imperative the test scores provide a valid and reliable reflection of the knowledge, skills, and abilities measured by the assessments. Unfortunately, the proclivity for individuals affected by the test results to engage in inappropriate behavior undoubtedly escalates as the stakes associated with the outcomes increase (Cohen & Wollack, 2006). It should come as no surprise that pressure created by the high-stakes nature of accountability testing is commonly cited as the motivation behind security violations (Kingston, 2013). To help combat the validity threat that test security violations pose, SEAs and LEAs must have policies and procedures in place to safeguard the integrity of assessment results and the decisions informed by them (National Council on Measurement in Education, 2012).

Although the previous paragraphs have all referred to “cheating,” we argue that it is important to make a distinction between behaviors that are nefarious, naïve, or somewhere in between. We reserve the word *cheating* to describe nefarious intent. We use the phrases *inappropriate behavior* and *security violations* as umbrella terms to describe any issue affecting test security whether nefarious or naïve. For educators, an understandable difficulty may arise in part because a considerable portion of educators’ normal instructional role is to help confused students achieve clarity, help them gain understanding, and provide scaffolding for student success. Although educators also typically introduce summative testing environments in their own classrooms, the change to an environment in which students are to independently demonstrate their

knowledge and skills may pose a difficulty for some educators. Clarity and training in the differences in expectations for educator behavior between the environments and requirements for compliance can reduce inappropriate, but partially or fully naïve activities to varying degrees.

Rather than improving test security, one could argue that the solution is to cease to place high stakes on standardized tests, because doing so increases the probability of perverse incentives to breach security (see, e.g., Stanford, 2013; Strauss, 2012). We hold another view. There are policies in many areas of public interest that create powerful incentives for inappropriate behavior. Any policy that includes sanctions and rewards based on outcomes is simultaneously a policy statement of desirable outcomes and an incentive to improve or to game the system based not on achieving the desirable outcomes but appearing to do so. We disagree that we should cease to make policy statements about desirable outcomes simply because doing so creates an incentive for inappropriate behavior. We argue instead that the appropriateness of the incentives be evaluated, and that mechanisms for deterring, detecting, and responding to inappropriate behaviors should be designed and implemented.

Therefore, this chapter focuses on the need for improving methods of preventing, detecting, and responding to test security violations in the context of next generation assessments and accountability whether such issues arise nefariously, intentionally, or somewhere in between.

THE IMPORTANCE AND COST OF MAINTAINING TEST SECURITY

Recognizing the importance of test security in the pervasive high-stakes testing climate, experts in the field of education and measurement have recently begun to focus on disseminating comprehensive test security recommendations for educational agencies (NCME, 2012; Olson & Fremer, 2013; Wollack & Fremer, 2013). However, adherence to comprehensive test security recommendations can be costly to for educational agencies, which have seen their overall funding decrease substantially since the 2008 recession (Oliff, Mai, & Leachman, 2012).

Implementation of these measures can require SEAs and LEAs to devote resources to new logistical procedures (e.g., locking and storage mechanisms), test administration information (e.g., seating charts), technology (e.g., scanners and software to accurately detect and record erasure marks; software to detect anomalous responses, similarity in responses, and other potential markers of online irregularities), staff trained in this technology, staff with the statistical knowledge to compute the often technically complex detection indices, and staff trained in conducting appropriate investigations of potential and confirmed security breaches.

These new demands come at the same time states are facing pressures to increase the quality of their assessments with such costly enhancements as the use of innovative, technology-enhanced items, performance tasks, and a larger number of additional test questions requiring students to construct their own responses. These pressures have come not only from constituents but also from the implementation of ESEA flexibility (U.S. Department of Education, 2013a) and the adoption of more rigorous content standards across the U.S. requiring more sophisticated assessments (Common Core State Standards Initiative, 2015; Heitin, 2014).

Already strapped by the resource-intensive task of meeting federal and state accountability demands, reductions in resources, and demands for improved assessment

quality, educational agencies typically lack the requisite resources to devote strictly toward test security measures. It is important that policy makers holding the purse recognize the need for additional resources devoted to maintaining test security if the results are to be used in increasingly high-stakes ways.

WHAT DOES THIS MEAN IN PRACTICE? AN EXEMPLARY RESPONSE

To illustrate the resources required to appropriately address test security violations facing educational agencies in the next generation accountability era, the following section provides a brief case study of what we view as an exemplary recent review and revision of test security policies and practices. The New York State Department of Education (NYSED) oversees various statewide educational assessment programs, including, but not limited to, the Grades 3 through 8 English language arts and mathematics assessments, Regents high school graduation exams, and English as a second language assessments. In total, NYSED is involved in the administration of more than 5,000,000 individual assessments in a given year (Greenberg, 2012a). Although many of these assessments have multiple uses, among the primary purposes is the use of student test scores in accountability. Thus, the assessments are typically high stakes for students, teachers, and/or school administrators. Prompted by the increasing stakes attached to testing in the accountability era and recognition of the additional security threats incurred by these stakes, a self-requested external review of NYSED test security processes found deficiencies in several critical areas (Greenberg, 2012b). The summative report offered critiques that have been consistently echoed as areas requiring improvements throughout the test security literature, (Cizek, 1999; Fremer & Ferrara, 2013; National Council on Measurement in Education, 2012) including

- lack of written policies, procedures, and quality control mechanisms to standardize and guide handling of test security matters;
- insufficient training and resources for those in charge of test security duties; and
- antiquated data systems and forensic methods to identify sophisticated cheating methods.

Following the external investigators recommendations, NYSED established a dedicated test security unit tasked with developing standardized test security policies, investigating irregularities, and instituting effective data tracking and forensics mechanisms (Sciocchetti, 2012). Creation of this unit required NYSED to first devote resources to hire seven new employees with experience in legal investigations and information technology. Moreover, the considerable overhaul of test security systems involved budgeting resources for aggressive pursuit of fraud investigations, renovation of data-tracking systems, and creation of thorough contemporary standards for test security practices, along with various other tasks. Because of the significant resources needed to create this new unit and implement new policies and procedures, New York likely presents a unique case.

Despite our view that the progress made by NYSED has been exemplary, establishment of this unit is not a panacea to all test security concerns. The diversity of components necessary to address test security concerns is considerable. Those best suited to develop strong legal policies differ from those best suited to implement data collection and tracking systems, which differ from those suited to conduct and report data forensic analyses. Furthermore, individuals specifically trained in educational test security

are rare and thus typically command high salaries given their unique skill sets. The New York test security unit, as directed, focused primarily on policy and legal investigative components of the process. Data forensics practices were still notably outdated and difficult for the unit to utilize effectively. Recognizing this remaining deficiency, NYSED requested and was provided an additional \$500,000 specifically toward analytics to detect test irregularities in both 2012 and 2013 (New York State Department of Education, 2013), which is currently being devoted to renewing data forensics practices.

Maintaining appropriate funding for such a unit to perform well over time will be a challenge. New ways to breach test security will be devised, and new methods of dealing with them will need to be developed, which may increase costs over time. And even with a dedicated unit, it is not possible to investigate all potential markers of irregularities. In nearly all cases, a unit must rely on random or targeted sampling to achieve its goals.

It is also important to note that test security work can be conducted internally, as in the NYSED example just described, or externally by contracting a company that specializes in this area. External contracting offers various benefits, such as obviating the need to hire test security specialists and the support of professionals who likely have a wealth of experience in educational test security. However, the benefits should be weighed against the potential costs, which can include less control over the security protocol, opaque proprietary methods, and substantial fees associated with the work. Each testing agency likely operates within its own unique set of circumstances that make either internal or external test security support a more appealing solution.

Needs in a Broader Arena

The confluence of increased motivation to cheat and general lack of available resources to devote toward security measures has created a difficult situation for agencies responsible for maintaining test security. Although resources may be scarce, test security must be given the necessary attention: without effective processes to deter and detect security violations, stakeholders will rightly lose confidence in the test scores used to support critical educational decisions. Acutely aware of the importance of test security, educational administrators and policy makers are relying on assessment and measurement experts to not only offer the most sophisticated techniques to identify irregularities but also provide practical guidance that can be implemented effectively within the constraints facing these agencies (Olson & Fremer, 2013). To address this goal, the remaining sections of this chapter first identify and synthesize security threats associated with specific testing contexts commonly used in educational testing. Next, we provide a comprehensive slate of potential actions agencies can employ to tighten test security and help safeguard the validity of assessment scores. We close by highlighting critical test security issues facing educational agencies transitioning toward next generation assessments. In general, this chapter attempts to provide efficient and practical recommendations for educational agencies and policy makers to maintain integrity of test scores in the accountability era.

SECURITY VULNERABILITIES IN CONTEMPORARY TESTING

In this section, we break down test security threats into categories of threats, categories of testing, and categories of sources of threats to assist in comprehensively describing threats and methods to address them.

Three basic types of accountability testing are currently being implemented and will likely continue to be implemented by the same agency in various combinations for the

foreseeable future. These are fixed-form paper-based testing, fixed-form online testing, and adaptive online testing. We also include performance tasks (PTs) as a basic form, although it can occur on all three forms previously mentioned because they come with their own security concerns if used as part of the assessment system. Each basic type comes with associated test security concerns.

Paper-and-pencil testing will likely remain in use for some time given the lack of technology at least in some schools. For example, in most states, if computers are mandated for testing, the state must pay for the devices above and beyond normal state funding to local districts. In addition, there will likely continue to be special circumstances requiring the availability of paper and pencil (e.g., schools serving a preponderance of Amish students, or paper-based testing as an accommodation). For at least three reasons, fixed-form online testing is likely to continue for some time as well: (1) because of the Partnership for Assessment of Readiness for College and Career (PARCC) decision to implement a fixed-form assessment (K-12 Center at Educational Testing Service, n.d.); (2) because of comparability concerns about moving away from fixed-form testing; and (3) due to the greater cost and effort involved in developing an adequate pool of items to implement adaptive testing. Even where adaptive testing is being implemented, such as in states adopting Smarter Balanced (K-12 Center at Educational Testing Service, n.d.) and states implementing their own adaptive assessments (Utah State Office of Education, n.d.), some subject areas may be presented in an adaptive manner and others in fixed form.

Threats to test security can be either naïve or nefarious and can come from five broad sources: test takers, educators formally involved in testing, and educators not formally involved in testing, top-level agencies responsible for assessment, and external parties. Following Martineau (2013), we also classify types of threats into the following categories:

- | | |
|-------------------|--|
| 1. Assistance | Non-test takers assisting test takers in variously inappropriate ways |
| 2. Copying | Test takers assisting each other in various inappropriate ways |
| 3. Exposure | Obtaining (or making available) specifics about what is on the test in various ways |
| 4. Gaming | Manipulating machine-scoring algorithms to achieve a high score without demonstrating necessary content knowledge and/or skill |
| 5. Identification | Falsifying the identity of test taker to improve scores (including erasing incorrect answers and replacing them with correct answers by someone other than the test taker) |
| 6. Materials | Students are allowed access to materials prohibited during testing |
| 7. Obstruction | Individuals refuse to cooperate with and/or deliberately obstruct an investigation |
| 8. Pressure | Directing subordinates to engage in inappropriate behavior or putting extreme pressure to perform on subordinates and/or students |
| 9. Proctoring | Not following test administration procedures during testing |
| 10. Selection | Inappropriately manipulating the population of test takers and/or manipulating data to improve aggregate outcomes. |

The difficulty in appropriately addressing test security can be seen in that the many factors (three types of testing plus performance tasks, two levels of intentionality, five sources of threats, and 10 types of threats) can combine in a large number of ways. In a

comprehensive strategy for safeguarding test security each threat needs to be anticipated, monitored for, and addressed if and when they occur. Thus, to adequately respond to threats to security, a reasonably comprehensive listing of specific threats is needed. To create such a compendium, we use as a foundation the threats identified in the CCSSO test security guidebook (Olson & Fremer, 2013), the NCME white paper (2012), a U.S. Government Accountability Office report on test security (2013), *Operational Best Practices for Statewide Large-scale Assessment Programs* (Council of Chief State School Officers & Association of Test Publishers, 2013), the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), and a report on a testing integrity symposium held by the U.S. Department of Education (USED, 2013b).

In addition to the threats identified in the documents listed above, we added information regarding specific threats to security on large-scale *online* educational accountability assessments compiled by Martineau (2013), who used as a foundation a document from the organization FairTest (n.d.) identifying over 50 ways that schools cheat on testing, based on compilations of reports from government and media sources. Martineau supplemented this foundation with interviews of strategically selected individuals with experience in (1) conducting data forensics on online testing, (2) administering online testing programs, and/or (3) administering a testing program during the transition from paper-and-pencil to CBT. The people interviewed were the following, who each gave permission to be named:

- John Fremer, the President of Caveon Test Security, a company that specializes in data forensics as applied to testing.
- Tony Alpert, the former Director of Assessment and Accountability for the Oregon Department of Education, where he was responsible for the implementation of computer adaptive assessment to fulfill the requirements of NCLB from 2005 through 2011. At the time of the interview, Alpert was the Chief Operating Officer of the Smarter Balanced Assessment Consortium.
- Carissa Miller, the former Deputy Superintendent responsible for assessment and accountability at the Idaho Department of Education. Miller oversaw the transition from paper and pencil assessment to online assessment. She is currently the Deputy Executive Director of the Council of Chief State School Officers.
- Judy Park, the Associate Superintendent responsible for assessment and accountability for the Utah Department of Education. She recently oversaw Utah's transition from paper and pencil to online assessment.
- Dirk Mattson, the former Director of Research and Assessment for the Minnesota Department of Education. He oversaw the transition to online assessment for Minnesota. At the time of the interview was the Executive Director of K-12 Multistate Assessment Programs for Educational Testing Service.
- Jennifer Dugan, Director of Statewide Testing for the Minnesota Department of Education, responsible for the current implementation of online assessment in Minnesota.

The interview included the following 12 questions allowing free form responses:

1. What are the threats to test security that an online fixed-form assessment poses?
2. What are the threats to test security that an online computer-adaptive assessment poses?

3. What unique threats to test security do constructed response test questions pose?
4. What unique threats to test security do performance tasks pose?
5. What unique threats to test security does artificial intelligence scoring pose?
6. What are student online cheating behaviors you are aware of?
7. What are teacher online cheating behaviors you are aware of?
8. What are administrator online cheating behaviors you are aware of?
9. What steps have you taken to address online test security?
10. What steps do you take to detect online cheating?
11. What is your greatest fear for online test security?
12. What question should I have asked that I did not ask?

These interviews provided information on recent, real-world experiences with threats to test security, cheating behavior, preventive measures, and detection measures in the transition to and implementation of large-scale online assessment. Separate questions regarding online fixed-form (FF) assessment and computer adaptive (CAT) assessment are included because PARCC and Smarter Balanced intend to implement fixed form and CAT assessment, respectively (PARCC, 2010; Smarter Balanced, 2010). The question regarding constructed response (CR) questions and performance tasks (PT) addressed PARCC and Smarter Balanced intent to include many of these types of tasks. Finally, the question regarding artificial intelligence (AI) scoring was included because both consortia indicate intent to use AI scoring to the degree feasible.

After compiling the lists of threats from the documents and interviews describe above, and as a final step in attempting to produce as comprehensive as possible a compendium of threats to test security, we (the current authors) supplemented this information with our own experiences.

Our compendium is presented as Table 15.1. In this table, we identify whether the threats apply to paper-based fixed-form testing, online fixed-form testing, online adaptive testing, and tests involving PTs (whether paper- or computer-based, fixed-form or adaptive). We also identify whether the threats can come from individual test takers, involved educators, other educators, outside parties, or the agency responsible for testing; and whether the threats could arise naively (all threats could arise nefariously). Last, we evaluate the threat in terms of its likely degree of threat to the validity of the score of any individual student (*validity*), potential scale of any single incident (*scale*), and likelihood of all incidents of a given type cumulatively affecting a substantial proportion of student scores (*likelihood*). These were rated subjectively, coded, and scored in the following manner:

Validity	Code	Score	Scale	Code	Score	Likelihood	Code	Score
Low	L	1	Individual Student	I	1	Very Low	VL	1
Moderate	M	2	Group of Students	G	2	Low	L	2
High	H	3	Classroom	C	3	Moderate	M	3
Critical	C	4	Building (school)	B	4	High	H	4
			District	D	5	Very High	VH	5
			Region	R	6			
			State	S	7			

Table 15.1 bears some explanation. Where a threat would be substantially (but not entirely) ameliorated by adding depth to the pool of content (e.g., a deep item pool for adaptive testing with strong exposure control, a deep pool of PTs with strong exposure control, a large number of forms randomly assigned to students) the threat is marked with an open circle (○ as compared to ●). We identify the validity impact as critical (C) where a score for a student could be expected to bear no resemblance to a valid score for that student. In terms of scoring the threats, we could identify no compelling reason to weight any of these three sets of ratings (validity, scale, and likelihood) more heavily than another, so to calculate an overall risk score for each threat we divided each score by the maximum possible score for that category to give each category a maximum risk score of one. Because larger threat scores on multiple categories pose a substantially greater risk than a large threat score in any one category, we multiplied (rather than summed) the scores from the three categories to derive a total score with a theoretical range from approximately zero¹ to one. We also did not have a strong basis for using any particular scoring scheme, so we used equidistant numbers for each change in level. We then rank ordered the threats based on the overall score to identify the most critical threats to address. Rank ordering the threats allows an agency to prioritize its resource expenditure to address the most pressing test security concerns.

Although the findings presented in Table 15.1 are useful, they are likely to need adjustment depending on the context of any given testing program. The evaluation and scoring we applied in Table 15.1 were based on our generic evaluation of threats across testing programs rather than for any specific testing program. Additional threats could be identified. Some threats may not apply. Other categories of evaluation (instead of or in addition to the three we chose) could be used. Different levels could be defined within the categories we used, and a different set of scores could be applied to those levels. Finally, different raters will give different subjective ratings. How each of these is operationalized would depend on the testing program.

To determine appropriate actions to take to address these security risks, the same set of documents (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Council of Chief State School Officers & Association of Test Publishers, 2013; Martineau, 2013; National Council on Measurement in Education, 2012; Olson & Fremer, 2013; U.S. Education Department, 2013; U.S. Government Accountability Office, 2013) was used to identify the various actions agencies can take to improve test security. We compiled recommended security actions from those documents, supplemented it with our own experiences, and present the compendium in Table 15.2, where the column headers directly correspond to the broad types of threats described in Table 15.1.

As with Table 15.1, we believe that what we present in Table 15.2 is useful, but will need adjustment depending on the context of any given testing program. The evaluation and scoring we applied in Table 15.1 were based on our generic evaluation of test security enhancement actions across testing programs rather than for any specific testing program. Additional actions could be identified. Some actions may not apply. A different rating scale could be used. Finally, different raters will give different subjective ratings. How each of these is operationalized would depend on the testing program.

The actions are grouped into several categories, their purpose is listed (whether it is to minimize naïve test security violations, deter breaches in security, detect breaches in security, and/or to respond to breaches in security. As do other resources (National Council on Measurement in Education, 2012; Olson & Fremer, 2013), we place a premium on those actions that deter test security breaches and minimize naïve breaches.

Table 15.1 Compendium of Threats to Test Security

Type	#	Description	Applies to	Sources	Threat Evaluation	Threat Score	Rank
Assistance							
Exposure	1	Helping students read/understand prompts and items	Paper fixed form	Involved Educators	H	G	68
	2	Helping students answer items	Online fixed form	Outside parties	H	C	57
	3	Placing helpful information on board	Online adaptive	Could be naive	M	C	78
	4	Giving inappropriate reminders during testing	Performance tasks	Testing agency	M	C	0.13
	5	Encouraging students to return to specific items/sections	Test-taker	Other educators	H	G	0.13
	6	Signalling students one or more answers are (in)correct	Involved Educators	Outside parties	M	C	78
	7	Exposing test content to later testers	Outside parties	Could be naive	M	C	0.13
	8	Enhancing or giving extra accommodations	Other educators	Testing agency	M	G	0.09
Copying							
Assistance	9	Answer sharing/copying by proximity	Paper fixed form	Involved Educators	H	G	68
	10	Answer sharing/copying by signal	Online fixed form	Outside parties	H	G	0.09
	11	Answer sharing/copying via electronic device	Online adaptive	Could be naive	M	G	86
	12	Students exposing test content to later testers	Performance tasks	Testing agency	C	C	89
	13	Proctors ignoring answer sharing/copying	Test-taker	Other educators	C	L	0.17
	14	Proctors facilitating answer sharing/copying	Involved Educators	Outside parties	C	L	68
	15	Small number of forms	Outside parties	Could be naive	H	S	0.75
	16	No dedicated breach form in the event of a critical breach	Involved Educators	Testing agency	C	VH	3
	17	No within-class spiraling of test forms	Outside parties	Other educators	H	S	1
							16

(Continued)

Table 15.1 (Continued)

Type	#	Description	Threat to Security	Applies to	Sources	Threat Evaluation	Threat Score	Rank
	18	Same test form administered on multiple days				H	S	0.45
	19	Long test window				H	S	0.45
	20	Shallow pool of test items				H	S	0.45
	21	Poor item exposure control				H	S	0.45
	22	Memorizing enough of upper end of pool to improve score				M	G	0.03
	23	Items placed on social media				H	S	0.6
	24	Insufficiently secure administration software				•	H	0.3
	25	Insufficiently secure administration hardware				•	H	0.3
	26	Insufficiently secure operating system				•	H	0.3
	27	Insufficiently secure technology hosting platform				•	H	0.3
	28	Paper booklets inappropriately obtained				•	H	0.26
	29	Copies of paper booklets made				•	H	0.26
	30	Unaccounted for secure paper materials				•	H	0.21
	31	Hacking client or vendor systems				•	H	0.6
	32	Hacking local systems				•	H	0.32
	33	Hackers expose live test form(s)				C	S	1
	34	Hackers expose entire item bank				H	S	0.6
	35	Remote desktop viewing applications				M	D	0.21
								64

Exposure, continued...

		Identification	Gaming	Exposure, continued...
36	Memorable test content	o	o	o M S M 0.3 42
37	Taking pictures of live test content	o	o	o M G M 0.09 89
38	Taking notes on live test content	o	o	o M C L 0.09 86
39	Social media sharing and discussion of live test content	o	o	o H G M 0.13 78
40	Targeting instruction toward live test content	o	o	o C R M 0.51 15
41	Shallow pool of performance tasks (PTs)	●	●	● H S M 0.45 16
42	Poor PT exposure control	o	o	o H S M 0.45 16
43	Multiple-day PTs	●	●	● M S M 0.3 42
44	Memorable PTs	o	o	o M S H 0.4 32
45	Administering PTs as a separate event	●	●	● L S VL 0.05 94
46	Administering classroom components as part of PTs	●	●	● H S M 0.45 16
47	Leaving paper materials in potentially exposed locations	●	●	● H D L 0.21 61
48	Not following chain of custody protocols	●	●	● M D L 0.14 76
49	Media in testing environment	●	●	● H S H 0.6 5
50	Other external parties in testing environment	●	●	● M B L 0.11 83
51	Insufficient training in chain of custody	●	●	● M D H 0.29 44
52	Insufficient training in storage and handling protocols	●	●	● M D H 0.29 44
53	Insufficient training in administration protocols	●	●	● M D H 0.29 44
54	Gaming the scoring algorithm (keywords, length)	●	●	● M I VL 0.01 96
55	Changing student scores (via hacking)	●	●	● C D L 0.29 44
56	Local scoring	●	●	● H S H 0.6 5
57	Insufficiently secure scoring software	●	●	● H S L 0.3 36
58	Insufficiently secure score database	●	●	● H S L 0.3 36
59	Changing student answers (erasing and replacing)	●	●	● H D H 0.43 24
60	Changing student answers (transfer to new answer doc)	●	●	● H D H 0.43 24
61	Changing student answers (before submission)	●	●	● H D H 0.43 24
62	Changing student answers (via hacking)	●	●	● H D L 0.21 61

(Continued)

Table 15.1 (Continued)

Type	#	Description	Threat to Security	Applies to	Sources	Threat Evaluation	Threat Score	Rank
						Likelihood ^a	Score	
						Validity ^a	Scale ^b	
						Could be naive		
						Testing agency		
						Outside parties		
						Other educators		
						Involved Educators		
						Test-taker		
						Performance tasks		
						Online adaptive		
						Online fixed form		
						Paper fixed form		
						MATERIALS		
						Obstruction		
	63	Filling in answers left blank by students			H	D	0.43	24
	64	Educators finishing online test for student			C	D	0.57	11
	65	Educators take test for student			C	D	0.29	44
	66	Key logging to gain access to student test event			C	D	0.29	44
	67	Educators log in as students to gain access to test content			H	D	0.54	14
	68	Switch student identifications			C	D	0.29	44
	69	Access to the internet during testing			M	B	0.11	83
	70	Access to prohibited materials (cheat sheets, calculators)			H	B	0.17	68
	71	Access to prohibited functions (spell check, thesaurus)			M	B	0.06	93
	72	Prohibited materials on walls (formulae, definitions)			M	B	0.17	74
	73	Prohibited materials on erasable boards			H	C	0.19	67
	74	Fabricating test security documentation			H	D	0.43	24
	75	Warning staff regarding security monitors			H	D	0.32	34
	76	Refusing security monitors immediate access as needed			C	D	0.43	30
	77	Refusing security monitors access as needed			C	D	0.29	44
	78	Lying to security monitors and/or investigators			C	D	0.29	44
	79	Disguising testing locations to prevent monitoring			C	B	0.11	83

	Selection	Proctoring	Pressure
80	Directing staff to cheat		
81	Educator proctoring his/her own class		
82	Educator proctoring the class of a relative or friend		
83	Collusion among shuffled classroom proctors		
84	Extreme pressure on administrators/teachers		
85	Extreme pressure on students		
86	Leave classroom unproctored		
87	Inattentive proctoring		
88	Too many students per proctor to be effective		
89	Insufficient local training for proctors		
90	Manipulate student grade levels to skip tests		
91	Exclude likely low-scorers from testing		
92	Fail to ship answer documents from likely low-scorers		
93	Fail to submit online tests from likely low-scorers		
94	Push likely low scorers out of school		
95	Falsify demographic data		
96	Urge likely low-scoring students to be absent		

a. Level of threat to validity of student scores (L = low, M = moderate, H = high, C = critical).

b. Potential scale of effects on validity (S = student, G = group of students, C = classroom, B = building/school, D = district, R = region, S = state).

c. Likelihood that cumulative events may effect a large number of students (VL = very low, L = low, M = moderate, H = high, VH = very high).

Table 15.2 Recommended Actions, Their Purposes, and Subjective Ratings of Effectiveness

Action	Purpose				Effectiveness in Dealing With...									
	Naïve	Deter	Detect	Respond	Assistance	Copying	Exposure	Gaming	Identification	Materials	Obstruction	Pressure	Proctoring	Selection
Certification	Certification that accommodations match IEP/504/ELL plans	•	•		H								H	
	Certification of training by trainers and trainees	•	•		H	L	H	H	H	H	H	M	H	H
	Certification of which proctor for which student for which test	•	•	•	H				H				H	
	Certification of examinee identity by appropriate party	•	•	•					H					
	Security/confidentiality/compliance forms by all agency staff	•	•	•				H						
	Security/confidentiality/compliance forms by all vendor staff	•	•	•				H						
	Security/confidentiality/compliance forms by involved educators	•	•	•	H	L	H	H	H	H	H	H	H	H
	In-person and electronic monitoring of training	•	•	•	M	M	M	M	M	M	M	M	M	M
Monitoring	In-person and electronic monitoring of preparation	•	•				H		H	L				
	In-person and electronic monitoring of chain of custody	•	•				L			H				
	In-person and electronic monitoring of test administration	•	•		M	M				M	M	M	M	M
	In-person and electronic monitoring of processing	•	•					M						
	Monitor maximum students per proctor	•	•	•								M		
	Monitor IEP/504/ELL accommodations	•	•	•			M							
	Random sample monitoring	•	•	•	M	M	M	M	M	M	M	M	M	M
	Targeted monitoring	•	•	•	M	M	M	M	M	M	M	M	M	M
Resources	Dedicated staff and/or contracted services to monitor security	•	•	•	M	L	H	H	L	L	M	L	L	H
	Budget for dealing with a security breach			•	M	M	M	M	M	M	M	M	M	M
	Detailed vendor security plan from item development through reporting	•	•	•	•			H						
	Detailed client security plan from item development through reporting	•	•	•	•			H						
	Qualified investigators on retainer or on staff			•	•	M	M	M	M	M	M	M	M	M
	Hotlines (phone, Internet, email) to report irregularities	•	•	•	•	M	M	M	M	M	M	M	M	M
	Monitoring for new threats to security			•		M	M	M	M	M	M	M	M	M
	Hotlines (phone, Internet, email) to ask about unforeseen issues	•	•	•	M	M	M	M	M	M	M	M	M	M
Depth	Deep item pool with strong exposure control	•	•	•	H	H	H							
	Deep PT pool with strong exposure control	•	•	•	H	H	H							
	PT classroom activities give topical background knowledge only ²			•	H	H	H							
	Large number of tasks randomly assigned to students within PTs ²			•	H	H	H							

Action	Purpose				Effectiveness in Dealing With...								
	Naive	Deter	Detect	Respond	Assistance	Copying	Exposure	Gaming	Identification	Materials	Obstruction	Pressure	Proctoring
Forensics	Large number of forms randomly assigned to students	•	•	•	H	H	H						
	Detailed, written plan for conducting and documenting data forensics		•	•	H	L	M	L	M				H
	Postmortem review of data forensics for analyses to add and/or remove		•	•	H	L	M	L	M				H
	Monitoring of internet for secure test content	•	•		L			M					
	Require IEP/504/ELL accommodations in advance and track use	•	•		L								
	Erasure analyses		•			M							
	Forensic analysis of nontested students		•										H
	Similarity analyses (including those using seating charts)		•		M	M		M					
	Person-fit analyses		•		M	M		M					
	Analysis of testing times, IP addresses, response time, test length		•		H	L	L		H				
Instructions	Improbable cohort score residual analyses		•		H				H				
	Logging of key paper-based administration events		•		M		L			L			
	Logging of key online test administration events		•		M			M					
	Monitor for missing secure materials		•					H					
	Requirements seating chart documentation		•			M							
	Detailed steps for monitoring chain of custody and storage	•	•				M						
	Detailed steps for returning materials	•	•	•				M					
	Detailed steps to prepare for testing	•	•				M						
	Detailed steps to remove prohibited materials/devices during testing	•	•				M		M				
	Detailed, scripted test administration manuals	•	•		M		M		M				
Policies	Prohibitions on sharing test content in scripted test admin manual	•	•			L	L						
	Authority to require cooperation with in-person monitoring	•	•	•	M	M	M		L	M	M	M	
	Authority to require compliance	•	•	•	H	H	H	H	H	H	H	H	H
	Authority to apply consequences for breaching security	•		•	H	H	H	H	H	H	H	H	H
	Authority to protect whistleblowers	•	•		M	M	M	H	M	M	M	M	H
	Clear delineation of consequences for noncompliance	•	•	•	H	H	H	H	H	H	H	H	H
	Consensus on security protocols across all users of a test	•	•	•			H						
	Clear definition of appropriate/inappropriate behavior for all roles	•	•	•	M	M	M	M	M	M	M	M	M

(Continued)

Table 15.2 (Continued)

Action	Purpose				Effectiveness in Dealing With...									
	Naïve	Deter	Detect	Respond	Assistance	Copying	Exposure	Gaming	Identification	Materials	Obstruction	Pressure	Proctoring	Selection
Requirements	Clear definition of allowable/nonallowable materials	•	•	•						H				
	Clear definition of vendor/client roles	•	•	•					H					
	Detailed protocols for addressing missing secure materials		•	•	•				M					
	Guidelines for appropriate use of results	•	•	•						L				
	Require testing of all eligible students	•	•						L		H			
	Rules for allowable and nonallowable responses to student questions	•	•				M							
	Rules for receipt, inventory, and return of secure testing materials	•	•	•				H						
	Rules for storing materials before, during, and after administration	•	•					H						
	Rules for keeping testing facilities secure	•	•				H							
	Rules for acceptable test administration facilities and rooms	•	•			L	M	M		L		H		
	Rules for visitors during testing	•	•				H				L			
	Limits on number of students per proctor	•	•				M	M						
	Students taking the same form take it on the same day and/or time	•	•	•		L	L	H						
	Physical obstruction (distance, dividers) to deter copying	•	•				M							
Scoring	Conflict of interest reporting (e.g., proctoring own class, friend's class)	•	•				H							
	Require amelioration of reported conflict of interest	•	•				H							
	Limits on test preparation	•	•							M				
	Distributed scoring (only outside district, no identifying information)	•	•	•					H					
	Monitor scoring against randomly seeded validity papers	•	•						H					
	Significant read-behind local scoring by professional scorers	•	•						M					
	Strong training and qualification protocols for local scoring	•	•	•					M					
	Random read-behind of machine-scored constructed responses		•						M	L				
Systems	Monitoring of ill-fitting machine-scored constructed responses		•						M	L				
	Audit (on sample basis) local technology systems for requirements	•	•	•				M	M	L				
	Define standards for security of local systems	•	•	•				M	M	L				
	Limit administration software to secure systems	•	•	•			H							

		Purpose	Effectiveness in Dealing With...																		
Action		Naïve Deter Detect Respond	Assistance Copying Exposure Gaming Identification Materials Obstruction Pressure Proctoring Selection																		
Systems, continued...	Limit hardware to secure systems	• • •	H																		
	Limit operating system to secure systems	• • •	H																		
	Lock down administration software and hardware	• • •	H																		
	Disable system during nonoperational hours	• •	H																		
	Lock students out after completion of a section	• •	H																		
	Monitor client systems for intrusion	• •	H M																		
	Monitor vendor systems for intrusions	• •	H M																		
	Place strong security on client systems	• •	H L																		
	Place strong security on technology hosting systems	• •	H L																		
	Place strong security on vendor systems	• •	H L																		
Training	Protocols for secure data transfer	• •	H L																		
	Detailed training for proctors/coordinators/teachers/administrators	• •	H M M M M M M M H M																		
	Require training for proctors/coordinators/teachers/administrators	• •	H H H H H H H H H H																		
	Include training on inappropriate behavior and consequences	• •	H H H H H H H H H H																		
Visibility	Include training on identifying answer copying/sharing	• • •	H																		
	Visibility of consequences for inappropriate behavior	• •	H H H H H H H H H H																		
	Visibility of monitoring	• •	H H H H H H H H H H																		
	Visibility of data forensics analyses	• •	H H H H H H H H H H																		
Investigation	Visibility of whistleblower reporting options	• •	H H H H H H H H H H																		
	Standardized documentation of (potential) security breaches	• •																			
	Detailed definition of evidence needed to initiate an investigation	•																			
	Detailed protocols for investigations	•																			
Recovery	Use of trained investigators	•																			
	Detailed definition of triggers for consequences	•																			
	Protocols for appealing a finding and/or consequence	•																			
	Dedicated breach form	•																			
	Detailed plan for action when a potential breach is identified	•																			
	Detailed plan for action when a breach is confirmed	•																			
	Detailed communication plans for potential or confirmed breach	•																			

In addition, we provide a subjective rating of the effectiveness of the various actions in dealing with various threats to test security. Again, we may not have gathered every possible action to take, and the effectiveness may be rated differently by different people for different assessment programs.

The descriptions of actions in Table 15.2 are admittedly brief. They will need to be fleshed out and operationalized. One such source with additional detail regarding specific actions is the Council of Chief State School Officers' and Association of Test Publishers' operational best practices guide (2013); the other chapters in this volume also provide additional detail on specific actions.

It is also possible to use Tables 15.1 and 15.2 in another way. A testing agency may wish to engage in its own evaluation of security threats, identify the threats it deems important, rate them, and develop responses to them based on the ability to pursue the actions that would have the greatest impact on their greatest threats. This approach recognizes that it is unlikely that any agency would be able to enact all actions, and may need to be thoughtful about which actions are pursued for the greatest benefit.

As can be seen in Table 15.2, there are many more actions that can be taken to avoid naïve test security breaches and to deter test security breaches than to detect and respond to them. This is entirely appropriate. Test security breaches have great potential to be costly or catastrophic. Whereas detection and rapid response are important to containing the effects of security breaches, we agree with other scholars (National Council on Measurement in Education, 2012; Olson & Fremer, 2013; United States Education Department, 2013), that the most effective way to reduce the possibility of test anomalies and/or test breaches is to educate test administrators, teachers, and their supervisors by providing high-quality, detailed training, manuals and/or online modules on test security so local education agencies can easily refer back to information if they have any questions.

Proactive guidance and instruction will provide the sponsoring agency with documentation that sets forth clear expectations as to how the test administration should be conducted. Reactive methods to detect and respond to security breaches remain important, but most security breaches can be prevented through proactive training.

Finally, although the work in completing Tables 15.1 and 15.2, and cross-referencing the tables to prioritize prevention, detection, and follow-up action is labor intensive, it will serve the sponsoring agency well as a proactive approach it can take to safeguarding test security and to protecting the agency, local agencies, educators, and students from unnecessary disruption and risk.

PROACTIVE TRANSITIONS TO NEXT GENERATION ASSESSMENTS

There are a few things that will be important for agencies moving toward next generation testing, with some specific issues applicable especially to states moving from paper-based testing to CBT. This section details a few items that sponsoring agencies should include in their review of test security threats and actions to prevent, deter, and follow up on test security incidents.

Some states have experience administering online assessments while others do not (Martineau & Dean, 2012). States making the transition will require specialized training in test security for online assessment. For example, New Jersey has always administered paper-based state assessments. Only 23.7% of test administrators surveyed in New Jersey had ever administered an assessment online before they administered the

Partnership for Assessments of Readiness for College and Careers (PARCC) field test in 2014, whereas 78.8% of the test administrators surveyed in Tennessee said they had administered a CBT before the PARCC field test (Sinclair, Deatz, Johnston-Fisher, Levinson, & Thacker, 2015). Given the disparities in experience with online assessment, trainings will need to be tailored to the computer-based test administration experience level of the sponsoring agency and the local educators that will participate in testing.

In addition, sponsoring agencies and local agencies that have administered CBTs likely will be using a different technology platform when administering next generation assessments, and as contracts expire, may transition platforms again. One of the first items sponsoring agencies should consider is creating a document that outlines previous state assessment policies and procedures compared to the new policy and procedures. In states like New Jersey, with limited experience in administering online assessments, educators were anxious about policy shifts, but to the surprise of many, the majority of the test administration policies and procedures that New Jersey had in place did not change when converting to a CBT. However, some of the mechanisms did change. Creating a document that clearly states the similarities and differences makes it easier for individuals to focus and train on the differences. Providing this documentation for agencies that have experience with CBT is also critical, particularly if they are changing test platforms or will be using multiple test platforms (e.g., one testing platform for Science assessments and another for their ELA and Math assessments). Such a document will likely ease anxiety and thus preempt many test security incidents.

In most cases, statewide assessments are administered by teachers or educators within the school building. The purpose of training these individuals on how to deliver the assessments is to provide background on what to expect during the administration of the assessment and to make the administration as automatic as possible without many judgment decisions during the course of the administration. However, with the transition to a new assessment and/or platform, the trainings may include new expectations for test administrators. To ensure that test administrators adhere to the new policies and procedures, sponsoring agencies should create a section in the test administrator manual or another easily accessible location that introduces a “What if?” section for testing administrators on how to deal with some of the common issues that may impact the validity of the student score. Ideally, a frequently asked question (FAQ) section on situations that may be considered a test breach and a solution to prevent such breaches. The FAQ document should be provided and understood by all individuals involved in the administration of the assessment.

One of the most significant changes for states that have traditionally administered paper-based assessments and are now administering CBTs are the roles and responsibilities of individuals involved in the test administration process. With CBT, the technology coordinator role emerges as an important aspect of the successful administration of statewide assessments. The technology coordinator role adds a layer of complexity to the test administration team. States must clearly redefine the roles of test administrators, technology coordinators, and district test coordinators. For example, test administrators, technology coordinators, and test coordinators may not fully understand the complexities of each other’s responsibilities, nor may they be aware of how individual decisions affect the work of the others. The roles and responsibilities must be clearly defined so everybody knows the expectations of each other during the test administration. Most likely, test administrators and district test coordinators will need to know the basic technology components in order for the technology coordinators to monitor the networks and be on call for any potential issues that may impact testing at a district and/or school level.

For security purposes, the roles and responsibilities must specifically detail what the technology coordinator can and cannot do once he or she enters a test administration environment. Technology coordinators must be involved in the trainings and sign the confidentiality agreements just like test administrators and others involved in the test administration. In most circumstances when the technology coordinator enters the testing environment, the individual will need to view the error code on the screen and perhaps use the mouse or keyboard to diagnose a problem. Therefore, it is critical that the technology coordinator understand how to navigate technology issues within an active testing environment, causing the least amount of disruption to other students.

CONCLUSIONS AND CRITICAL ISSUES TO CONSIDER IN RESPONDING TO TEST SECURITY INCIDENTS

Responding to test security incidents can present significant risks to a sponsoring agency, local agencies, educators, and students. The act of investigation could impact test validity, and is likely to affect the work of administrators, teachers, and students. If it becomes known that an investigation is under way, it can throw a local education agency, school, or classroom into turmoil, negatively affecting the education of students. Reaching a conclusion that a breach has occurred (with or without intent) can have employment consequences for educators and life-changing effects for students (e.g., invalidation of a score important to the student's future).

There are risks for anyone involved in testing from inappropriately launching, conducting, or reaching conclusions in an investigation. Therefore, it is imperative that sponsoring agencies have clarity regarding their authority in requiring compliance with test security protocols, investigating potential security breaches, and responding to potential or confirmed test security breaches. It is also imperative that the sponsoring agency has clear protocols for evaluating quantitative and qualitative data regarding a potential breach, clearly defined data-based triggers for launching an investigation, clear protocols for conducting an investigation, clearly defined standards for reaching a conclusion that a security breach has occurred, clearly defined standards for determining intent, clearly defined data-based triggers for applying consequences in the event of confirmed security breaches, and clearly defined procedures for appealing the findings of an investigation and application of consequences.

Our use of the word *data* is intentionally broad. There may be cases in which upon first blush the primary data (such as erasure analyses or improbable gains analyses) suggest a clear security breach and clear intent. However, there may be reasonable explanations (e.g., teacher forgot to pass out answer documents to the right students and therefore had to instruct students to erase their answers and hand back their answer documents, and then redistribute the answer documents to the right students) that constitute a mistake in administration, but not a security breach, and would not be characterized as ill-intentioned. Our use of the word *data* includes secondary data obtained from an investigation.

Although there may be instances in which it is justified to conclude that a breach occurred (with or without intent) based solely on primary data, we argue that it is always appropriate to ask whether there are any plausible explanations that do not constitute a breach or do not constitute intent, regardless of how rare such an event might be. Asking and following up on such a question may protect the sponsoring agency, local agency, educators, and students from unintended and undeserved consequences.

Another important consideration is that when a test security concern is triggered, the analysis of multiple measures will strengthen the case. It may be necessary to wait

to act on a data trigger until another year of data is available to avoid reaching a tenuous conclusion, or to triangulate across multiple data sources within a single administration, or both. Sponsoring agencies should be more comfortable with stronger action if they continue to witness anomalies in testing data or test breaches at the same school or district over multiple years and/or on multiple criteria.

An important point is that the burden of proof in determining that a potential breach should be investigated, that a breach has occurred, and that consequences are warranted is the responsibility of the sponsoring agency. Therefore, the sponsoring agency must be able to justify its actions with clear documentation that it proactively engaged involved parties regarding appropriate behavior, had authority to take action, and did so on a clearly defensible and documented basis.

Finally, although the authority to take action is imperative, flexibility is also important. Whatever mechanism exists to give authority to the sponsoring agency to require compliance with testing procedures and security protections must also allow the agency flexibility to respond to new threats to security, new types of security breaches, and loopholes in existing requirements. Legislation, regulations, rules, or policies defining the authority must not so tightly bind the sponsoring agency that it cannot nimbly respond to changes in the test security landscape.

NOTES

1. The minimum possible theoretical score is 1/140, or 0.007143.
2. PTs organized around a theme, with classroom activities focusing on the background knowledge about the theme (to level the playing field for all students), and random assignment of many sets of items/tasks by student such that direct instruction on all content a student may see for a given PT will be prohibitive.

REFERENCES

- Achieve, Inc. (2015). *DCI arrangement of standards*. Retrieved February 16, 2015, from Next Generation Science Standards: www.nextgenscience.org/search-standards-dci
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Cohen, A. A., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355–386). Westport, CT: American Council on Education/Praeger.
- Common Core State Standards Initiative. (2015). *Standards in your state*. Retrieved February 10, 2015, from Common Core State Standards Initiative: www.corestandards.org/standards-in-your-state/
- Conley, D. T., Drummond, K. V., de Gonzalez, A., Seburn, M., Stout, O., & Rooseboom, J. (2011). *Lining up: The relationship between the common core state standards and five sets of comparison standards*. Eugene, OR: Educational Policy Improvement Center. Retrieved February 16, 2015, from www.epiconline.org/publications/documents/LiningUp-FullReport_2011.pdf
- Council of Chief State School Officers & Association of Test Publishers. (2013). *Operational best practices for statewide large-scale assessment programs*. Washington, DC: Authors.
- Duncan, A. (2011, July 19). Despite cheating scandals, testing and teaching are not at odds. *The Washington Post*. Retrieved from www.washingtonpost.com
- FairTest. (n.d.). *50+ ways schools “cheat” on testing: manipulating high-stakes exam scores for political gain*. Retrieved April 14, 2013, from FairTest.org: <http://fairtest.org/sites/default/files/Cheating-50WaysSchools-ManipulateTestScores.pdf>
- Florida Department of Education. (n.d.). *Student progression*. Retrieved February 9, 2015, from Florida Department of Education: wwwfldoe.org/academics/standards/student-progression
- Fremer, J. J., & Ferrara, S. (2013). Security in large scale, paper and pencil testing. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 11–17). New York, NY: Routledge.

- Georgia Department of Education. (n.d.). *Georgia high school graduation tests (GHSGT)*. Retrieved February 9, 2015, from Georgia Department of Education: www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/GHSGT.aspx
- Goertz, M., & Duffy, M. (2003). Mapping the landscape of high-stakes testing and accountability programs. *Theory Into Practice*, 42(1), 4–11.
- Graham, K. A. (2014, September 27). 2 more Phila. educators charged in cheating probe. *philly.com*. Retrieved February 16, 2015, from http://articles.philly.com/2014-09-27/news/54357171_1_philadelphia-school-district-pssas-bok-high-schools
- Greenberg, H. M. (2012a). *Ensuring the integrity of the New York state testing program* [PowerPoint slides]. Retrieved February 6, 2015, from www.regents.nysesd.gov/meetings/2012Meetings/March2012/TestIntegrityPowerPoint.pdf
- Greenberg, H. M. (2012b). *Review of the New York State Education Department's ("NYSED") processes and procedures for handling and responding to reports of alleged irregularities in the administration and scoring of state assessments*. Retrieved February 6, 2015, from www.regents.nysesd.gov/meetings/2012Meetings/March2012/test-integrity-report.pdf
- Haymes, S. (2013). *A comparison of national and Alaska common core education standards*. Juneau, AK: Alaska State Legislature Legislative Research Services. Retrieved February 16, 2015, from www.nealaska.org/sites/default/files/Attachment%20to%20Q&A.pdf
- Heitin, L. (2014). *Next generation science standards: Which states adopted and when?* Retrieved February 10, 2015, from Education Week: http://blogs.edweek.org/edweek/curriculum/2014/08/next_generation_science_standa.html?r=1252920820
- Hull, J. (2013). *Trends in teacher evaluation: How states are measuring teacher performance*. National School Boards Association Center for Public Education. Retrieved February 9, 2015, from www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/Trends-in-Teacher-Evaluation-At-A-Glance/Trends-in-Teacher-Evaluation-Full-Report-PDF.pdf
- Jacob, B. A., & Levitt, S. D. (2003). *Rotten apples: An investigation of the prevalence and predictors of teacher cheating*. Cambridge, MA: National Bureau of Economic Research.
- Judd, A. (2012, September 22). School test cheating thrives while investigations languish: Allegations of test tampering often get limited scrutiny. *Atlanta Journal-Constitution*. Retrieved from www.ajc.com
- K-12 Center at Educational Testing Service. (n.d.). *Key similarities and differences between the comprehensive assessment consortia*. Retrieved February 2, 2015, from K-12 Center at Educational Testing Service: http://k12center.org/rsc/pdf/key_consoritia_flyer_hr.pdf
- Kingston, N. M. (2013). Educational testing case studies. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 299–311). New York, NY: Routledge.
- Martineau, J. A. (2013). *Test security in the context of developing computer-based common core assessments*. Paper presentation at the annual meeting of the National Council on Measurement in Education. San Francisco, CA.
- Martineau, J. A., & Dean, V. J. (2012). A state perspective on enhancing assessment and accountability systems through systemic implementation of technology. In R. L. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessments: Recent history and predictions for the future* (pp. 55–78). Charlotte, NC: Information Age Publishing.
- Martinez, A., & Anderson, L. (2016, April 27). 5 educators arrested in EPISD scheme. *El Paso Times*. Retrieved from www.elpasotimes.com
- Mills, C. N., & Steffen, M. (2002). The GRE computer adaptive test: Operational issues. In W. J. van der Linden & G. A. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 75–99). New York, NY: Kluwer Academic Publishers.
- National Council on Measurement in Education. (2012). *Testing and data integrity in the administration of statewide student assessment programs*. Madison, WI: Author.
- National Governors Association & Council of Chief State School Officers. (2010). *The standards*. Retrieved March 24, 2013, from Common Core State Standards Initiative: www.corestandards.org/the-standards
- New York State Department of Education. (2010). *100.5 Diploma requirements*. Retrieved March 10, 2015, from New York State Department of Education: www.p12.nysesd.gov/part100/pages/1005.html
- New York State Department of Education. (2013). *Commissioner King announced new test security website: Latest step toward protecting integrity of state assessments [press release]*. Retrieved February 6, 2015, from New York State Department of Education: www.nysesd.gov/news/2015/commissioner-king-announces-new-test-security-website
- No Child Left Behind Act of 2001. (n.d.). Pub. L. No. 108–110.
- Ohio Department of Education. (n.d.). *Student promotion and the third grade reading guarantee*. Retrieved February 9, 2015, from Ohio Department of Education: <http://education.ohio.gov/Topics/Early-Learning/Third-Grade-Reading-Guarantee/Third-Grade-Reading-Guarantee-District-Resources/Student-Promotion-and-the-Third-Grade-Reading-Guar>

- Oliff, P., Mai, C., & Leachman, M. (2012). *New school year brings more cuts in state funding for schools*. Washington, DC: Center on Budget and Policy Priorities. Retrieved February 6, 2015, from www.cbpp.org/cms/?fa=view&id=3825
- Olson, J., & Fremer, J. J. (2013). *TILSA test security guidebook: Preventing, detecting, and investigating test security irregularities*. Washington, DC: Council of Chief State School Officers.
- PARCC. (2010). *Race to the top assessment program: Application for new grants*. Retrieved April 14, 2013, from U.S. Department of Education: www2.ed.gov/programs/racetothetop-assessment/rtta2010parcc.pdf
- Pell, M. B. (2012, September 30). More cheating scandals inevitable, as states can't ensure test integrity: National education policy built on test scores is undermined. *Atlanta Journal-Constitution*. Retrieved from www.ajc.com
- Perry, J., & Vogell, H. (2008, December 14). Surge in CRCT results raises 'big red flag'. *Atlanta Journal-Constitution*. Retrieved from www.ajc.com
- Perry, J., Vogell, H., Judd, A., & Pell, M. B. (2012, March 25). Cheating our children: Suspicious school test scores across the nation. *Atlanta Journal-Constitution*. Retrieved from www.ajc.com
- Poggio, J., & McJunkin, L. (2012). History, current practice, perspectives and what the future holds for computer based assessment in K-12 education. In R. L. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessments: Recent history and predictions for the future* (pp. 2–54). Charlotte, NC: Information Age Publishing.
- Sanchez, C. (2013, April 10). *El Paso schools cheating scandal: Who's accountable?* Retrieved April 14, 2013, from National Public Radio: www.npr.org/2013/04/10/176784631/el-paso-schools-cheating-scandal-probes-officials-accountability
- Schaeffer, G. A., Reese, C. M., Steffen, M., McKinley, R. L., & Mills, C. N. (1993). *Field test of a computer-based GRE general test*. Princeton, NJ: Educational Testing Service.
- Sciocchetti, T. E. (2012). *Ensuring the integrity of the New York state testing program: Test security unit update*. Retrieved February 6, 2015, from www.highered.nysed.gov/tsei/documents/tsufinal1ppt_1.pdf
- Sinclair, A., Deatz, R., Johnston-Fisher, J., Levinson, H., & Thacker, A. (2015). *Findings from the quality of test administrations investigations: PARCC field tests*. Unpublished manuscript.
- Smarter Balanced. (2010). *Smarter balanced assessment consortium application*. Retrieved March 27, 2013, from U.S. Department of Education: www2.ed.gov/programs/racetothetop-assessment/rtta2010smarterbalanced.pdf
- Stanford, J. (2013, May 23). The cheating will continue until morale improves. *Huffington Post*. Retrieved February 16, 2015, from www.huffingtonpost.com/jason-stanford/standardized-test-cheating_b_3325239.html
- Strauss, V. (2012, February 17). How to stop cheating on standardized tests. *The Washington Post*. Retrieved February 16, 2015, from www.washingtonpost.com/blogs/answer-sheet/post/how-to-stop-cheating-on-standardized-tests/2012/02/16/gIQAPF0nIR_blog.html
- Toppo, G. (2013, April 11). Memo warns of rampant cheating in D.D. public schools. *USA Today*. Retrieved February 16, 2015, from www.usatoday.com/story/news/nation/2013/04/11/memo-washington-dc-schools-cheating/2074473/
- Toppo, G., Amos, D., Gillum, J., & Upton, J. (2011, March 17). When test scores seem too good to believe. *USA Today*. Retrieved February 16, 2015, from http://usatoday30.usatoday.com/news/education/2011-03-06-school-testing_N.htm
- U.S. Department of Education. (2010a). School improvement grants: American recovery and reinvestment act of 2009 (ARRA): Title I of the elementary and secondary education Act of 1965, as amended (ESEA). *Federal Register*, 75(13), 3375–3383.
- U.S. Department of Education. (2010b). *U.S. Secretary of Education Duncan announces winners of competition to improve student assessments*. Retrieved March 24, 2013, from U.S. Department of Education: www.ed.gov/news/press-releases/us-secretary-education-duncan-announces-winners-competition-improve-student-asse
- U.S. Department of Education. (2012). *ESEA flexibility request*. Retrieved April 24, 2013, from U.S. Department of Education: www.ed.gov/esea/flexibility/documents/esea-flexibility-request.doc
- U.S. Department of Education. (2013a). *ESEA flexibility*. Retrieved April 14, 2013, from U.S. Department of Education: www.ed.gov/esea/flexibility/requests
- U.S. Department of Education. (2013b). *Testing integrity symposium: Issues and recommendations for best practice*. Washington, DC: USED Institute of Education Sciences National Center for Education Statistics. Retrieved February 17, 2015, from National Center for Education Statistics: <http://nces.ed.gov/pubs2013/2013454.pdf>
- U.S. Government Accountability Office. (2013). *K-12 education: States' test security policies and procedures varied*. Washington, DC: Author. Retrieved February 17, 2015, from www.gao.gov/assets/660/654721.pdf
- Utah State Office of Education. (n.d.). *Utah's comprehensive accountability system*. Retrieved February 12, 2015, from Utah State Office of Education: www.schools.utah.gov/assessment/Adaptive-Assessment-System/SAGEUCASBrochureWeb.aspx
- Wollack, J. A., & Fremer, J. J. (Eds.). (2013). *Handbook of test security*. New York, NY: Routledge.

16

ESTABLISHING BASELINE DATA FOR INCIDENTS OF MISCONDUCT IN THE NEXTGEN ASSESSMENT ENVIRONMENT

Deborah J. Harris and Chi-Yu Huang

INTRODUCTION

In high-stakes testing environments, such as when a student's promotion to the next grade or whether a teacher retains his or her position depends on the result of a testing event, incidents of assessment impropriety that may affect the validity of test scores are a concern. Statistical indices and procedures designed to flag or help detect instances of potential misconduct are prevalent, and the development of new techniques and methodologies appears to be flourishing (see, for example, the programs of the Annual Conferences on Test Security). Although it is acknowledged that statistical methods in and of themselves cannot prove some type of irregularity occurred (see, for example, Cizek, 1999; Harris & Schoenig, 2013), they can be strong tools in identifying situations deserving a closer look and protecting the integrity of test scores and the decisions based on them.

Multiple indices and procedures have been developed, based on various methodologies and measurement theories. Because incidents of misconduct vary (for example, sending in a surrogate test taker differs from a test administrator changing an examinee's responses after testing, which also differs from an examinee visiting a chat room where an item from an upcoming test has been shared from a test preparation company), the techniques to identify likely misconduct are also varied. For some of these situations there may be an assumption that the computed indices follow a particular distribution, allowing one to identify a value that is unusual, provided one specifies what unusual means (i.e., 1 in 1,000 or 1 in 1,000,000). However, for some of the procedures and statistics computed there is no basis to assume a particular distribution, which makes determining what is unusual, and therefore worth additional investigation, more difficult. This may be compounded in situations involving next generation assessments, where there is also little historical knowledge of examinee behavior on the assessments. Baseline data can provide a comparative basis, without distributional assumptions.

Much of the previous literature on detecting testing anomalies has focused on single format assessments, especially multiple-choice tests. Today's next generation assessments are often mixed format, composed of both multiple choice and other item formats, such as various types of constructed-response items, technology-enhanced items, or other newly developed formats. Studies on, for example, examinee copying on technology-enhanced items, or what is excessive overlap on constructed response text between two examinees, are sparse. Li, Huang, and Harris (2014) presented some analyses that might be used with mixed-format tests, at both an individual examinee and an aggregate level. An important question to ask when investigating potential anomalies is: "What criteria to use to determine if further investigation is warranted?" That is, is there sufficient indication that an anomaly may have occurred to warrant continued investigation? This is especially important with both novel assessment designs and new indices because there is typically no literature or best or typical practices to draw on. In their study, Li et al. used criteria of beyond two standard errors. However, even a criterion based on standard errors requires either distributional assumptions or a comparative group (such as baseline data) as a reference.

Several studies in the incidents of misconduct literature have looked at empirically developed baseline data to compare various statistics computed on suspect examinees to those computed on a larger group, making use of both simulated and real data. For example, Tracy, Lee, and Albanese (2014) describe using baseline data for a high-stakes licensure assessment to help interpret indices created to help detect examinee copying. They cite creating the baseline data using examinees who tested in different physical locations and therefore who presumably could not have copied from each other. They use the percentage of examinee pairs in the baseline data as a way to categorize how unusual a result for a suspect pair of examinees is. For example, the authors cite a value that less than 1% of the baseline data meets or exceeds as "moderately unusual." The authors also mention that assuming a normal distribution did not work well in terms of estimating the false positive rates. Hanson, Harris, and Brennan (1987) also found that the false positive rates based on theoretical assumptions did not agree well with the false positive rates produced using baseline, and recommended the latter be used whenever possible. Other studies making use of some type of comparison, or baseline, information include Jacob and Levitt (2003), Geranpayeh (2014), and Kanneganti, Fry, Gupta, and van der Linden (2014).

One of the challenges in using statistics to detect anomalies that might be evidence of potential misconduct is that there are not obvious criteria for evaluating whether an index value is unusual enough to warrant further consideration. Baseline data can often be helpful in this regard, not that it results in universally accepted values for determining whether to further investigate but that it can provide an indication of, at least in a comparative sense, what may be considered an atypical value. This may be especially helpful in situations involving novel assessments, where there is little historical performance to draw on for interpretation. For example, over time, how examinees initially score, who are subsequently found to have copied or to have used a surrogate, may be accumulated and used to help inform what to screen for in future administrations of the assessment.

In the current climate of next generation assessments there is an increasing emphasis on monitoring growth over time, novel item types, computer delivery, and on making high-stakes decisions at the individual student and aggregate (teacher, school, or district) level on the basis of assessment results. The emphasis on actionable decisions based on test scores makes identifying incidents of misconduct (in essence, avoiding

making decisions on invalid and inaccurate data) important. The fact that the items, administration, delivery, and scoring of assessments on which student scores are based is changing and more novel, makes identifying incidents of potential misconduct from natural variations more challenging.

BASELINE DATA

Statistical probability can be a strong indicator of how unusual an event is, but in its raw form, it may at times be misleading, or at the very least, not the best criterion to use in investigating incidents of potential misconduct. To expand on Holland (1996), chance alone will likely result in every examinee agreeing “to some positive degree” with other examinees. Recall that our goal is to determine whether an incident of misconduct has likely occurred, not to try to prove something may or may not have happened. For example, consider the relatively straightforward situation where one examinee is suspected of copying from another examinee, based on their proximity to each other during testing and on the similarity of their responses. If we take a situation where the assessment is a four-option multiple-choice test, we could assume the chance of both examinees selecting the same response on any given item was $\frac{1}{4} \times \frac{1}{4}$. However, we know this is not the case if both examinees happen to know the correct answer to an item. And it likely is not true if both examinees do not know the correct answer. Examinees often have partial knowledge, which may lead them to a popular distractor or away from others. Baseline data can be constructed to take into account distractors’ varying level of appeal by pairing students who are unlikely to have copied (such as two examinees testing in different rooms or different states) and looking at how often they have chosen the same distractors on various items.

Baseline data provide a rich context in which to interpret results. It also serves as a way to account for variables and conditions that can’t really be accounted for, or can’t be accounted for easily. When looking at how similar two sets of examinee responses are, it is typically not known what the tendency is for examinees who answer C to Item 7 to answer D to Item 8, or how likely a student getting a score of 3 on a constructed response item is likely to get a score of 2 on the next technology enhanced item. Baseline data can take this tendency into account, albeit indirectly. Consider two pairs of examinees, one pair that has 40 item responses in common and the other pair that has 24 item responses in common. Examining additional information may be helpful, such as how many of those common responses are incorrect; examining such pairs in relation to baseline data may be particularly helpful.

For example, suppose that the pair of examinees with 40 responses in common had two identical incorrect responses, and the pair of examinees who had 24 responses in common had 18 identical incorrect items. Although both the larger number of identical responses, and the larger number of identical incorrect responses, seem more suspicious, the absence of a criterion makes it difficult to determine how unusual these values are. A comparison to baseline information, however, can provide some context to aid in determining if either pair warrants further investigation.

Figure 16.1 provides baseline data for pairs of examinees, based on the number of identical responses and the number of identical incorrect responses they share. It can be seen from the data that the first pair (i.e., the pair that shares 40 overall responses in common) does not look atypical compared to the baseline data, whereas the second pair (i.e., the pair with only 24 overall responses in common) looks more atypical. Baseline data in this case provide additional context to evaluate the overall number

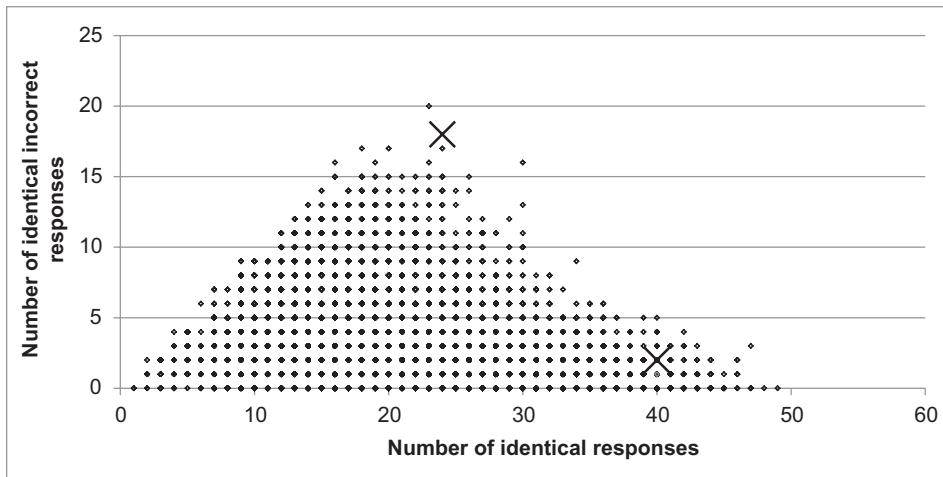


Figure 16.1 Bivariate Plot of Number of Identical Response and Identical Incorrect Responses of Pairs of Examinees

of item responses in common, as well as how the number of incorrect responses in common given a particular number of overall common responses may be interpreted as typical or atypical. In addition, the comparison to baseline data, such as shown in Figure 16.1, can also provide a visual representation of how unusual a value for a suspect pair is compared to other pairs of examinees. Such a visual representation may be easier for those evaluating the evidence and determining whether or not to pursue and investigation.

There are different ways that baseline data can be created. For example, the baseline data could be derived from a sample of the total population of examinees; or it could be based on all examinees; or by creating examinee pairs who could not have copied from each other (such as pairing two examinees who had tested in different physical locations); or by using only pairs of examinees with similar total scores; or by comparing all examinees within a particular location (e.g., school district), whether they tested in the same room or not, to control for having exposure to a common curriculum.

One of the benefits of using baseline data is that it is not necessary to make any parametric or other statistical assumptions. If a clear description of the data included in the baseline creation is provided, the user (e.g., test sponsor, agency, psychometrician, or arbitrator) can make the determination of the relevance of baseline data in a particular instance. Additionally, part of the evidence in pursuing a possible incident of alleged misconduct is noting whether something unusual happened, such as two examinees having an extreme number of identical responses, or too-similar item latencies. It is hard to know if a value is unusual or too similar; thus, it is important to examine comparisons with other data to understand the range of values typically obtained by most examinees.

With most triggers or indices of alleged misconduct, criteria for evaluating the unusualness evidence or statistical values need to be developed. One way to establish thresholds for unusualness could be to look at known cases of misconduct and known benign cases (for example, confessed copiers versus examinees who tested in different states or different administrations) and see what values of appropriate indices appear to separate the two groups. The practical disadvantage of this approach is

finding a large enough sample of known cheaters on which to compute the various indices, especially on new next generation assessments. Having comparative baseline data increases the likelihood of the values not being over or under interpreted and provides a context in which to present information if the decision is made to pursue a suspected cheating incident. For example, in the two pairs of examinees described previously, it might intuitively seem that two examinees with 40 common responses is much more unlikely than a pair with only 24 item responses in common, but looking at both the total number of common responses and the number of incorrect common responses, as well as comparing those values to the baseline data, would suggest that the second pair, rather than the first, would be the pair to further investigate if one chose to pursue either of them. As stated earlier, baseline data can be refined, or made more relevant, if more examinee information is collected, such as where the suspected examinees received their relevant training or based on their overall ability (either based on total test score, or on other variables, such as grade point average or previous test scores), or whether they employed the same tutor for test preparation. Hanson et al. (1987) did not find enough improvement with using conditional baseline data (where the data was matched on examinee ability ranges) to warrant its use in the situations they studied, however, so whether the additional effort involved in using conditional baselines data is worthwhile may be context specific. In situations involving novel indices and/or next generation assessments where the literature does not provide much guidance, it is prudent to examine conditional baseline data to determine empirically whether conditional comparisons provides similar or additional (confirmatory or contradictory) information.

ESTABLISHING BASELINE DATA

Establishing baseline data requires knowing what type of incident is suspected (see Harris & Schoenig, 2013, for a discussion of different triggers). For some types of misconduct, such as copying, it is relatively straightforward to create baseline data. For others, such as suspicion of a surrogate, it can be more challenging, as variations in handwriting, or degree of similarity with a photo, can be more difficult to amass than looking at the number of identical incorrect responses a pair of examinees has in common. Regardless of the index, however, the basic premise of using baseline data remains the same: the more the suspected examinee(s) differ from other examinees in their index value, the more support there is for something unusual having occurred. It is important to keep in mind that the goal of such investigations is to decide if an index value warrants or supports further investigation, and not make a definitive determination if an incident of misconduct occurred.

Baseline data can be created by computing the same indices on the group of examinees as one is computing on the suspect examinee(s). However, the choice or definition of the group is a primary issue. One way to compute baseline data would be to include only examinees or examinee pairs or groups that are not suspected of misconduct, or where the likelihood of misconduct is low. For example, a classroom with an external observer or auditor may be considered to be unlikely to have committed certain types of group irregularities. Similarly, two examinees who test in different test locations may be thought unlikely to have colluded on their responses, or an examinee testing with an advisor as a proctor may be thought less likely to be a surrogate than when an examinee tests at a site where he or she is unknown. In a sense, baseline data based on these types of groups could be considered as null groups, as it is thought the

index values computed on these groups represents normal variation, uncontaminated by including values based on misconduct.

Another option is to create baseline data on examinees in general, which may or may not include examinees who committed other incidents of misconduct (that is, there may be incidents of the same or different type of irregularities, but they will not be the same incidents and examinees under investigation). This type of baseline data is most typical of the baseline data used for initial screening, such as looking at the average number of erasures within each classroom to see if any classrooms look aberrant, where any classroom with suspected misconduct involving erasures would be included in the baseline data.

In addition, decisions around alleged incidents of misconduct often appropriately hinge on multiple sources of information. When computing baseline information on several indices, the question arises of what the relationship should be between the various sets of data. Because indices are often not independent, given that most analyses tend to be based on examinee-provided information related to the same testing event (e.g., latencies, answer choices, total scores), there is some basis for computing baseline data for all the indices of interest on the same group. This approach may be easier to interpret, especially to arbitrators and examinees, than using different, even nonoverlapping baseline data. Examinees with several index values that appear to be unusual compared to a single set of baseline examinees may provide a more coherent picture for interpretation than a situation where the comparison group needs to be redefined for each analysis. In addition, examinees with only one extreme value may also be easier to interpret, either as a benign event (some examinee has to be most extreme even in situations with no misconduct) or as information narrowing in on what exactly the alleged misconduct was likely to be. For example, a group of test takers with an extreme number of identical incorrect responses but a typical erasure analyses may have been aided by a test administrator who announced answers during the test administration rather than tampered with student answer documents after the test administration.

Looking at customized, or more relevant, subsets of examinees to create baseline data is also appropriate in many situations. For example, to address a claim that the reason two examinees had a large number of identical incorrect responses or produced similar essay responses was because they had the same training, either in school or a preparation class, baseline data based on only examinees who attended the same school, or had the same teachers, or enrolled in the same test preparation course, may be more relevant. Baseline data might also be created for each individual test form for some indices, such as in cases where the differential desirability of the incorrect options affects how likely two examinees are to independently choose the same incorrect response, but not in other cases, such as average erasures, which may be fairly consistent across grade levels, test forms or examinee cohorts. The former would likely be more dependent on particular items, whereas average erasures might be similar enough across test forms in some cases to have a single set of baseline data across forms.

In all cases where a comparison to baseline data is made, the key determinant of how typical or suspicious an outcome is, is how different the suspect examinee(s) index value compares to the baseline values. In any data set, some value will be most extreme. The judgment of whether to pursue an incident or not should consider the totality of evidence, both supporting and not supporting an allegation of misconduct, given the limited recourses most organizations have for these types of investigations, and the potential stress and harm of pursuing allegations that have little demonstrated merit (see Hanson et al., 1987; Olson & Fremer, 2013). Whether a value is atypical compared to baseline data, as well as how atypical that value is, helps inform the decision.

HOW THE DISPLAY OF BASELINE DATA FACILITATES INTERPRETATION

It should be stated at the onset that we are not recommending techniques to slant interpretation, such as adjusting the scale of the axis to make a value look more or less extreme. Rather, we recommend the use of data display (whether in tables, lists, figures, plots, or other graphics) to help interpret index values in relation to baseline data. Whereas some of the indices we compute to look at allegations of misconduct are straightforward (e.g., the number of erasures on an examinee's answer document or how many answers two examinees who sat next to each other have in common), other indices (e.g., person fit or the number of identical incorrect in the longest string of identical responses) are more difficult to conceptualize, particularly for an arbitrator, administrator, or other entity charged with making the decision of whether further investigation of an incident of suspected misconduct is warranted.

Comparing a value to baseline data provides a context in which to interpret the results, but only if comparison is easy to comprehend for those needing to make that determination. Probabilities may be difficult to understand for lay people, but seeing an "X" for a suspect pair of examinees that is far more extreme than any of the baseline data can be more intuitive to comprehend and evaluate. Therefore it is essential that those creating the baseline data choose appropriate comparative data and display results in a way that maintains the integrity of the comparison and does not distort it. The two graphs in Figures 16.2a and 16.2b display the percentage of erasures occurring by item for an exam with 25 multiple-choice items. The first plot displays the items by sequence (first item on the test, second item on the test, and so on until the last item on the test); the second plot shows the same information ordered by item difficulty (most difficult item first, through the easiest item, which appears last). Displaying items in two orders may provide insights into how erasures tend to occur. It perhaps is not surprising for more erasures to be observed at the end of a test, due to possible speediness and because more difficult items are often at the end of an exam. That does not

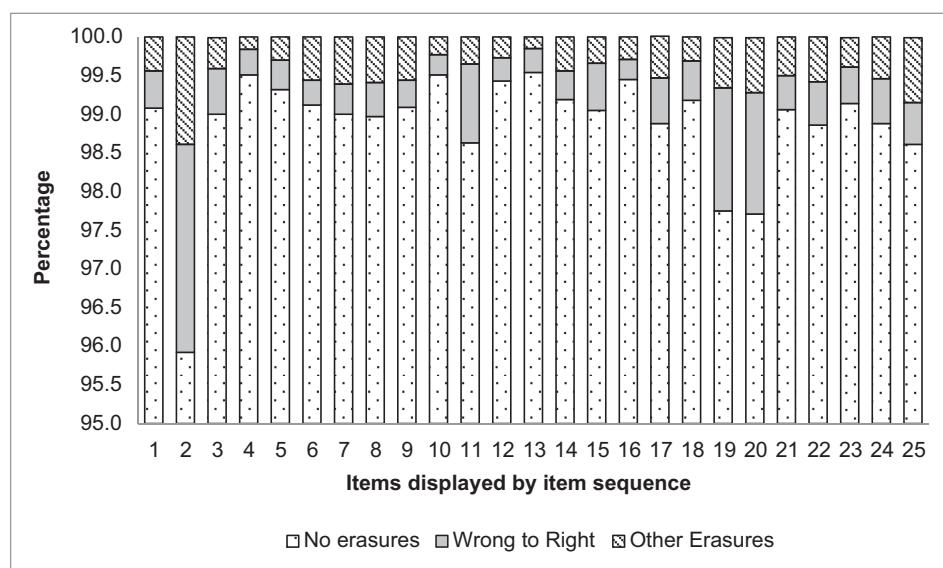


Figure 16.2a Percentage of Erasures Ordered by Item Sequence

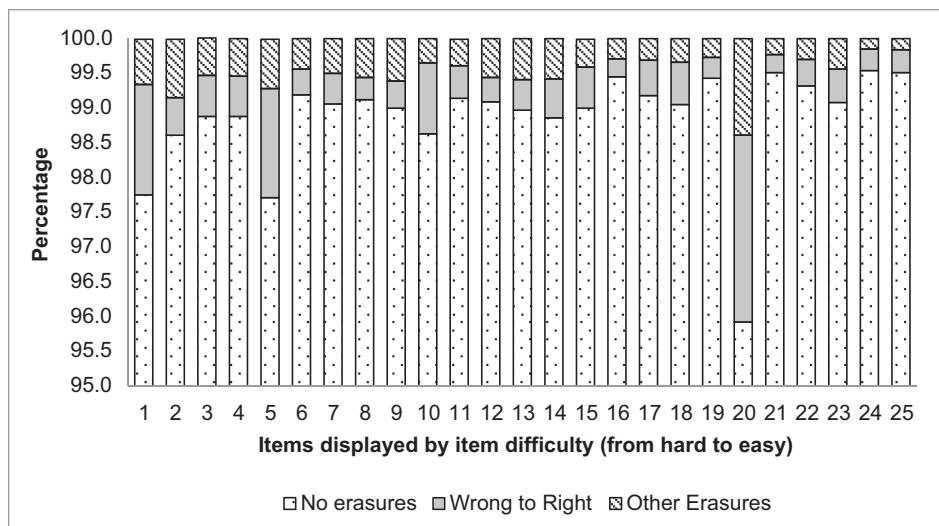


Figure 16.2b Percentage of Erasures Ordered by Item Difficulty

appear to be the case, however, for the data from the 25-item test represented in the figure. Looking at an examinee's pattern of erasures as well as wrong-to-right answer changes in relation to different baseline data (in this case, item information ordered in different ways) can provide additional information into how typical the examinee's behavior is. This can be especially helpful in cases where there are scrambled versions of a test, and seeing the items both ordered by position and by difficulty, as well as base form order, may provide additional insights.

Allegations of misconduct rarely rest on a single piece of evidence. Indices used to investigate misconduct often look at multiple variables, such as identical item responses and identical incorrect item responses. Computing baseline data on multiple variables can be informative. Consider a situation where an assessment consists of both multiple-choice and constructed-response items. In addition to looking at total information, looking at information by item type is recommended. For example, Figure 16.3 shows examinee latency information for the multiple-choice items and the constructed-response items on an online, next-generation, summative assessment. It can be seen that whereas the majority of examinees have a similar pattern, there are a smattering of outliers, including two quite extreme cases. If either of these two examinees had been suspected of, say, collusion, the fact that their latencies seem atypical might help inform if they should be considered the source or the copier. Of course, with a time stamp for every keystroke on a computer-based examination, this determination could be much more definitive, although not all testing programs collect this information. Displaying latency information by item type as a bivariate plot may lead to different insights than displaying the information for multiple-choice, constructed-response, and technology-enhanced items by item sequence, as is done in Figure 16.4. The outliers in Figure 16.3 may more easily lead one to question why the latencies on the constructed-response items were atypical: Was it because the examinee had memorized a response and did not need time to compose one, or because they left the response field blank, or some other reason? Also, looking at the latencies by subject (e.g., perhaps the examinee did better on algebra-related items than those measuring

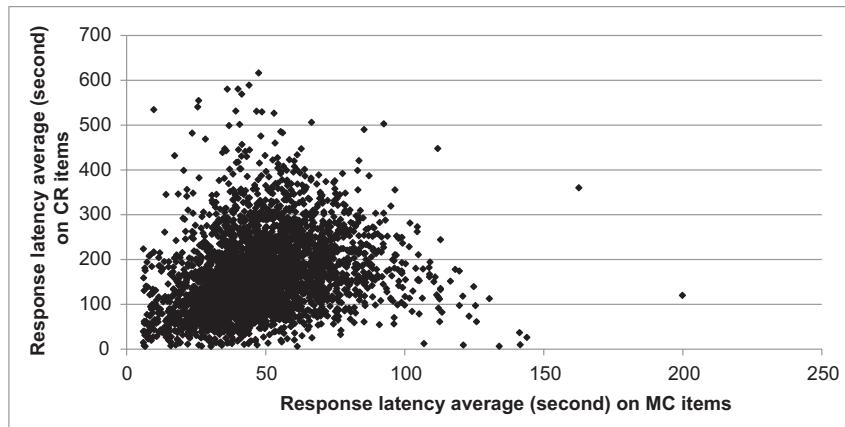


Figure 16.3 Individual Item Latency by Item Type (CR = constructed response, MC = multiple choice.)

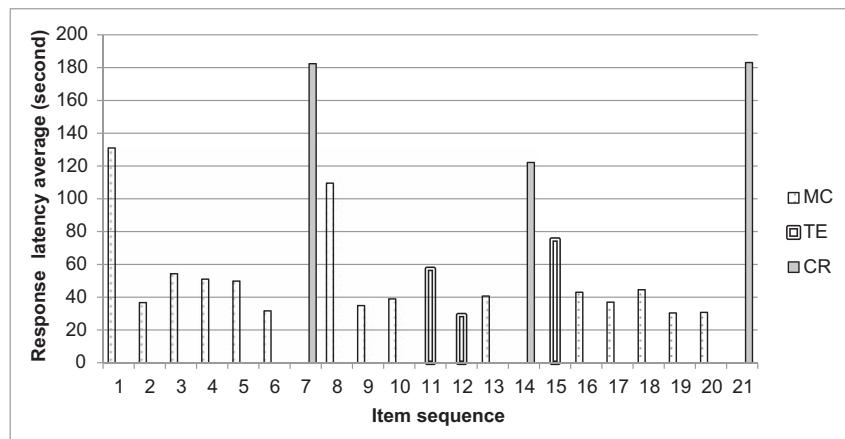


Figure 16.4 Item Type Latency by Item Sequence (MC = multiple choice, TE = technology enhanced, CR = constructed response.)

geometry knowledge and skill) can also be informative. The main point is that displaying baseline data in different ways, grouping on different variables, combining different variables, facilitates insights and interpretation of what may, or may not, be values suggestive of an incident of misconduct. See Harris, Huang, and Dunnington (2013) for additional examples of how visualization can aid in interpretation.

Examples

Below are two examples of how baseline data can facilitate interpretation of index values for further consideration.

Example 1: Possible Preexposure

One frequent concern in high-stakes testing is that items may be compromised prior to test administration. This may occur because of preexposure of, for example, paper test

booklets, or from reuse of particular items. A school administrator who feels strong pressure for students to do well may provide copies of a test early for review by a teacher or students, or a teacher who wants to prepare his or her students for the testing experience may not realize that using an old test booklet from the administration two years ago as a practice test exposes students to linking items to the current year's test. In addition, students or sites that test later than other sites may find out about test content through social media and other means. There may also be blatant misconduct during an assessment when, for example, a supervisor announces one or more answers during an administration, or changes student responses after an administration but before submitting student responses for scoring. Regardless of whether there is nefarious intent, the result is the same: students may have preexposure to test content, which may affect the validity of their scores as a measure of achievement.

The example below illustrates a test site where misconduct was suspected on a test form that contained both new (never before administered) items and items that were considered secure, but had been used in a previous administration (repeated or linking items). A group consisting of testing sites and examinees that were not suspected of engaging in misconduct but who were otherwise similar to the suspect, or target, group was created and used to develop baseline data. The assumption was that, if an incident of misconduct of the type suspected had occurred, the students in the target group should score higher on the reused items, and lower on the new items (whether they were operational or field test items). Figure 16.5 shows both groups of examinees. On the horizontal axis are the raw scores of each examinee on the reused items that were suspected of being compromised in the target group. On the vertical axis are the raw scores of each examinee on the new, or not-suspected-of-being compromised, items. For the baseline data, the pattern is clear: students who do well on the reused items also do well on the new items. Throughout the range of ability (that is, students who score high or score low on the reused items), examinees tend to score similarly on the two sets of items. However for the target group, there is a different pattern for a cluster of students. These students perform very well on the reused items, but very poorly on the new items. Comparing the suspected test site to a baseline group provides a clear visual

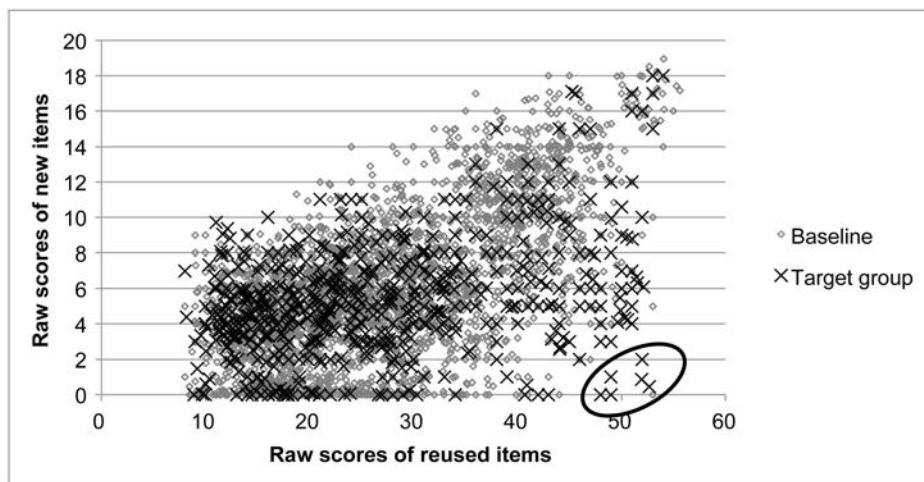


Figure 16.5 Total Scores on New and Reused Items

of a cluster of examinees that should be investigated further. For instance, did these students all take the exam under the same test supervisor? Do they belong on the same particular sports team or school club? Are they a social group? The graphic alone does not prove that an incident of misconduct occurred, but it provides strong evidence that there is reason to pursue an investigation, given the starkly different pattern of the suspect group from the baseline data. The comparison also shows that many of the target group performed like the baseline group, which suggests that, if there was an incident of misconduct at the site, it may only involve a subset of examinees. Comparing to a baseline provides a criterion for what looks unusual, but it is not an exact criterion, and other information, including the trigger than made one investigate this target group initially, needs to be considered. The figure seems consistent with some level of breach around previously used items; however, if the trigger was related to, for example, surrogate testing, the support is less clear.

Example 2: Item Latencies

With increased use of computer-based testing for next-generation assessments, the use of latency information, typically operationalized as the amount of time an item is on screen during an administration, has increased. The amount of time an examinee takes to respond to one item and move on to the next has been used as an indicator of whether the student is doing his or her own work. For example, if a test supervisor announced how to respond to an item to a class, the amount of time spent on an item might have no relation to how hard or easy the item was. It is acknowledged that there are many aspects that influence latency (e.g., the difficulty of the item, whether a student tries to work out an item or just omits or guesses if it seem too difficult, how much time is remaining in the testing session), but looking at item latencies and patterns of item latencies can be informative in investigating for possible incidents of misconduct.

In considering the latency pattern of a given examinee, baseline data can help determine if a pattern seems aberrant. For example, one might assume that examinees spend more time on difficult items than on easy items. Indeed, that seems to be the case for the total group data shown in Figure 16.6. An examinee with a different pattern

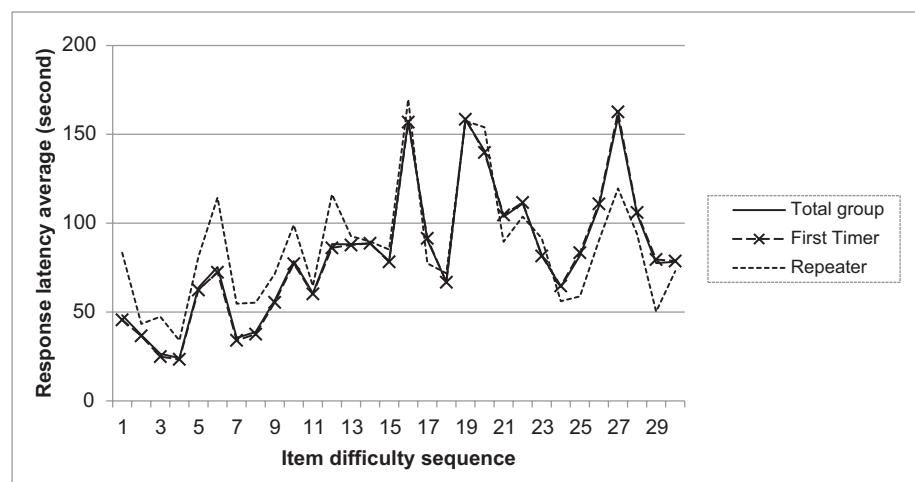


Figure 16.6 Item Latency by Item by Examinee Status

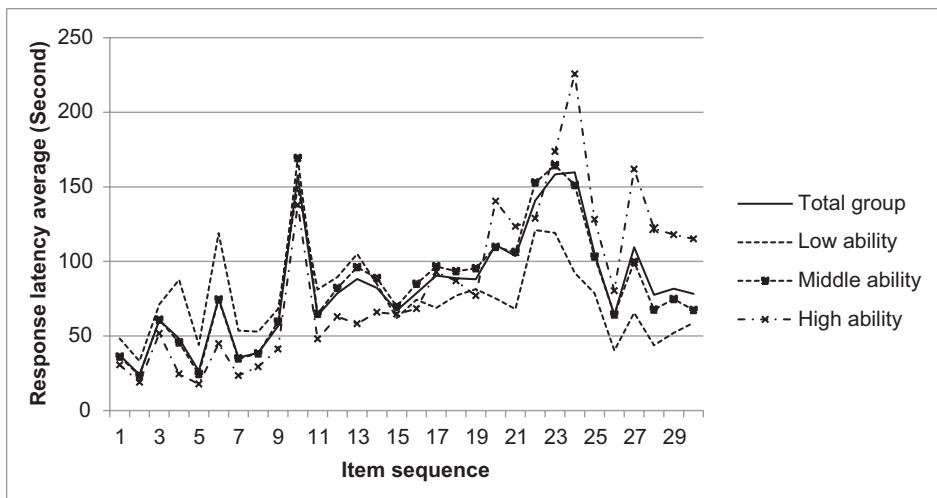


Figure 16.7 Item Latency by Item by Examinee Ability

might be deemed unusual by the comparison. However, if the examinee in question was a repeater examinee, who had taken the exam on at least one occasion prior to the administration in question, his or her latency pattern may differ from the total group, but still be typical for similar (i.e., repeater) examinees. This provides support for ensuring the baseline data are relevant to the context.

Additional breakdowns of latency data can be seen in Figure 16.7. Although sometimes finer levels or conditional baseline data may not provide much additional information (see, for example, Hanson et al., 1987), such analyses may be helpful in disputing a defense that the baseline data were not sufficiently like the potential defendant to provide accurate information for a comparison. For example, a plot like Figure 16.7 can demonstrate that even conditional on examinees of similar ability based on total score or some other metric, a particular item latency pattern was unusual.

BEST PRACTICES IN CREATING BASELINE DATA

As we have indicated previously, it should be noted that when baseline data are created and used, even extreme values compared to the baseline information are not proof that an act of misconduct occurred. As Olson and Fremer (2013, p. 29) remind us, “there will be some substantial differences even when all rules are followed scrupulously.”

The value in any comparison relies on the comparison criterion, making the selection of the baseline data of prime importance. To the extent it is practical and defensible, it is worth investigating customizing the baseline group to a specific context, particularly if preparing information for further investigation. For example, there may be a general baseline set of data to screen for, say, an unusual latency pattern or overlap of words used in an essay, with a second level of investigation involving comparison to baseline data specific to that particular essay, or to examinees who scored in a similar score range or who had taken a similar training course.

How to decide whether a value is extreme is not always obvious. An “X” marking an examinee pair far from all of the baseline data can present a compelling picture, but

as mentioned earlier, there is wide variation in some values even when incidents of misconduct do not occur, so it is quite likely there will be instances of overlap when an incident did occur. One can set an arbitrary cutoff, such as the top 5% or 1% of cases, or look for values that exceed a certain number of standard deviations from the average value. Or one can develop more complex or conjunctive rules, such as above a certain threshold (say, top 5%) on multiple index values and above a more stringent threshold (say, .5%) on at least one value.

Simulated data (data generated to fit a particular model) may also be useful in helping to set baseline data criteria, but the process is inherently limited because the intricacies and interdependencies of how real examinees perform are difficult if not impossible to generate. Looking at how index values of examinees who confess to misconduct compare to baseline data can be helpful (as in, would a particular comparison to baseline data with a particular threshold value have identified the examinee), but the number of examinees who confess may be too small to make this a practical approach. Manipulating real data can be helpful in situations where it is appropriate. For example, randomly selecting two examinees and forcing the copying of answers may be helpful in trying to determine a threshold of what looks extreme compared to different baseline data groups. However, trying to manipulate a surrogate situation may be a bit more challenging.

Another decision that needs to be made is the size of the group comprising the baseline data. Whereas more data is usually preferable to less, there are situations where the amount of available data is overwhelming. For example, if there are 70,000 examinees administered an assessment on a particular administration, and a group of 250 examinees at a particular test site are being investigated, it is possible that a better baseline comparison than using all 70,000 would be to randomly select a smaller subset of examinees (which would be easier to discern in a plot as well). This subset of data could be truly random or stratified to ensure it shared certain characteristics proportionally with the total group or it could be selected using, say, the same status (repeater, first timer) or score ranges as the investigated examinee(s).

Suspected incidents of misconduct that apply to groups of examinees can be examined through comparison on the individual level (see, for example, Figure 16.5, where some but not all of the suspected group were identified as having unusual values), or as a group. Creating group baseline data is also possible, such as based on the average number of erasures at a school level for an administration of a third grade summative math test. Because school achievement often correlates with various demographic variables (e.g., school size, location, percent of free and reduced lunch students), using schools or groups with similar characteristics can add to the perceived credibility of the comparisons.

As is familiar to anyone studying the literature on incidents of misconduct, everything seems to be context specific. Assumptions that work well with one assessment program don't work as well with another. Different types of possible misconduct are of primary concern in different situations. For example, an isolated event of one third grader copying off another's paper is probably less of concern in the arena of large-scale state-wide assessment than a teacher altering answer sheets before sending them in for scoring. However, on a licensure exam, one examinee copying off another may be of great concern. Similarly, how to compute the most appropriate baseline data can also vary. The overarching principles are to make the baseline data as relevant and appropriate as possible to the comparisons and inferences one wishes to make, and to not overinterpret the results of the comparison.

CONCLUSIONS

We hope that this chapter has provided the practitioner with both an appreciation for the role baseline data can play in the investigation of incidents of misconduct and some useful advice that can be used in the creation, display, and use of baseline data in practice. As new methods of investigation and new indices continue to be developed, baseline data can continue to serve as a useful tool in the interpretation of the values of the indices. We stress that there are always outliers; that is, one or more values will always be extreme. But by using multiple indices, and by using known examples of known (perhaps contrived) misconduct, baseline data can be a strong tool for identifying incidents of misconduct and avoiding misidentifying nonincidents.

Baseline data also provides a powerful vehicle for displaying results, visually displaying the discrepancy between the examinee of interest's values of a statistic with values from a large group of examinees. Baseline data can be customized to a particular examinee situation, and thus help address defenses such as "we have similar answers because we studied together" or "we took the same test prep course."

As the contents of this book demonstrate, there are many aspects to consider when investigating alleged incidents of misconducts. Using baseline data is one way to provide criterion information to help inform whether an alleged incident should be pursued or not. It can be used as part of an initial screening (as in, "Does this classroom have an unusual number of erasures compared to other classrooms?"), or to provide additional information once a determination to proceed has been made, or both. The use of baseline data is extremely flexible, and can rest on few or no assumptions, depending on how it is created. For example, a distribution of the average number of erasures per classroom can be made, with no assumption of whether some of these classrooms have been affected by misconduct. Similarly, distributions of the number of wrong answers examinees have in common can be made by matching pairs from different testing rooms who presumably could not have copied from each other.

Statistical indices, such as fit indices, the number of wrong-to-right erasures, and the amount of time an examinee spends in responding to an item, are part of the totality of evidence used in determining if an incident of misconduct has occurred during a test administration. However, it is often difficult to determine when a value is indicative of aberrance. Having baseline data to compare values to both increases the likelihood of the values not being over- or underinterpreted and provides a context in which to present information if the decision is made to pursue an incident of suspected misconduct.

REFERENCES

- Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum.
- Geranpayeh, A. (2014). *Answer changing patterns in computer-based tests*. Paper presented at Conference on Test Security, Iowa City, IA.
- Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying* (RR No. 87-15). Iowa City, IA: ACT.
- Harris, D. J., & Schoenig, R. R. (2013). Conducting investigations of misconduct. In J. A. Wollack & J. J. Fremer (eds.), *Handbook of test security* (pp. 201–219). New York: Taylor & Francis.
- Harris, D. J., Huang, C.-Y., & Dunnington, R. (2013). *Establishing baseline data for incidents of misconduct in the next generation assessment environment*. Paper presented at 2nd Annual Conference on Statistical Detection of Potential Test Fraud, Madison, WI.
- Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: statistical theory and empirical support*. (ETS Technical Report No. 96-4). Princeton, NJ: Educational Testing Service.

- Jacob, B. A., & Levitt, S. D. (2003). *Catching cheating teachers: the result of an unusual experiment in implementing theory*. (NBER Working Paper 9414). Cambridge, MA: National Bureau of Economic Research.
- Kanneganti, R., Fry, R., Gupta, L., & van der Linden, W. J. (2014). *Forensic erasure analysis on optical mark recognition documents*. Presentation presented at Conference on Test Security, Iowa City, IA.
- Li, X., Huang, C.-Y., & Harris, D. J. (2014). *Examining individual and cluster test irregularities in mixed-format testing*. Presentation presented at Conference on Test Security, Iowa City, IA.
- Olson, J. F., & Fremer, J. (2013). *TILSA test security guidebook, preventing, detecting, and investigating test security irregularities*. Washington, DC: Council of Chief State School Officers.
- Tracy, C., Lee, S. Y., & Albanese, M. (2014). *A comparison of similarity indexes for detecting answer copying on the MBE*. Paper presented at Conference on Test Security, Iowa City, IA.

17

VISUAL DISPLAYS OF TEST FRAUD DATA

Brett P. Foley

INTRODUCTION

As the stakes associated with tests increase, so too does the pressure to gain an unfair advantage through fraudulent means. Testing professionals have come to view test security and fraud detection as integral components in evaluating the validity of the intended uses and interpretations of test scores (Adams, 2014). Because of the importance of the issue, a great deal of research has been done on methods of fraud detection (see, for example, Kingston & Clark, 2014; Wollack & Fremer, 2013; other chapters in this volume). Many of the developed methods for test fraud detection depend on sophisticated quantitative techniques. However, the effectiveness of these methods will be limited if the information they contain cannot be effectively conveyed to stakeholders (e.g., administrators of testing programs, examinees, licensure boards) and other consumers of test security investigation results (e.g., the public, judges/juries, ethics panels). Many users of test fraud data likely will not have the technical expertise to fully understand the underlying mathematics of many of the quantitative methods for fraud detection that may be implemented. However, we believe that those involved with test fraud investigations have an ethical obligation to present results in ways that are accessible to the consumers of these investigations. To that end, research has shown that one way to reduce the cognitive burden associated with the interpretation of quantitative data is through the effective use of visual displays (e.g., Pastor & Finney, 2013).

The purpose of this chapter is twofold: (1) to present an explanation of characteristics of effective visual displays based on recommendations from the literature and (2) to present examples of visual displays that might work well for the types of quantitative metrics used in test fraud analyses. Rather than introducing specific quantitative methods for fraud detection, this chapter focuses on design principles and selecting visual displays to effectively convey the sometimes complex results of test fraud analyses.

CHARACTERISTICS OF EFFECTIVE VISUAL DISPLAYS

There is a broad body of literature addressing characteristics of effective visual displays. Several visual display scholars have put forth recommendations for choosing and

creating effective displays (e.g., Bertin, 1983; Cleveland, 1985; Few, 2012; Lane & Sáñ-dor, 2009; Mayer, 2013; Tufte, 2001; Wainer, 1997, 2005; Wong, 2010). Although these scholars sometimes provide conflicting guidance, several general design principles can be synthesized from the literature (Foley, 2015):

- Choose displays that are appropriate for the data.
- Aim for simplicity.
- Integrate text, numbers, and figures.
- Highlight what is important.
- Do not intentionally mislead.
- Pair the design with the audience.

In the following sections, each of these principles will be explored in the context of test fraud analysis. Examples are provided to show how these principles can be applied in practical situations.

Choose Displays That Are Appropriate for the Data

The choice of visual display can influence both the visual impact and the richness of data that can be shared (Wong, 2013). Before design and style decisions can be made, it is necessary to select an appropriate display type. There are a multitude of visual displays that can be employed for summarizing data, with countless variations of each of type of display. As such, a comprehensive discussion of all available display types is beyond the scope of this chapter. However, there are some broad guidelines that can serve as a basis for a starting point. One primary decision is the choice between using a table or graph. Tables are most useful when it is necessary to present specific values and when a significant amount of detail is necessary for completeness. Tables can also serve as effective formats for visual organizers. For example, Table 17.1 shows an excerpt from

Table 17.1 An Excerpt From a Table Used as a Visual Organizer Summarizing Test Security Issues and Related Information

Security Issue	Primary Threat			Data Source	Example Detection Methods
	IP	Validity	Financial		
Proxy test taking		X		Test scores	Unusual score gains on retest
				Biometrics	Palm Vein Fingerprint
Harvesting exam content	X	X		Test center incident reports	Manual review of CCTV video
				Specific (public) item content	Digital watermarking Web monitoring
Item compromise	X	X		Item-level data	Unusual response patterns
				Misfitting items	Item fit and drift analyses
				Brain-dump sites	Web monitoring & similarity analysis
Item pool compromise	X	X	X	Brain-dump sites	Web monitoring & similarity analysis
Answer key compromise	X	X	X	Brain-dump sites	Web monitoring & similarity analysis

Note. IP = intellectual property.

a visual organizer that might be useful for an organization developing a security plan for an examination. The organizer shows potential security issues, the type of threat they represent (e.g., loss of intellectual property), data sources that could be used for identifying the threat, and the specific fraud detection methods that could be brought to bear.

Graphs are more effective for showing trends and relationships. Graph choice will depend on the type of variables to be displayed and the type of relationship that is to be shown. For example, scatterplots are effective for showing the relationship between quantitative variables; bar graphs are useful for showing comparisons of a quantitative variable across several values of a categorical variable; histograms and dot plots are useful for showing the distribution of a quantitative variable; pie charts show the relationships between parts of a whole. In some situations, it can also be effective to combine multiple display types into a single visual display (e.g., a line graph overlaid on a scatterplot, a table and graph integrated into a single display). For additional guidance for choosing between tables and graphs and for selecting the appropriate graph type for the type of relationship that is to be featured, see Few (2012, pp. 39–52, 309–310) and Gillan, Wickens, Hollands, and Carswell (1998, pp. 29–31). Additionally, Harris (1999) provides a highly detailed taxonomy of display types, with extensive explanations and illustrations.

A corollary of choosing the appropriate display for the data is choosing the appropriate data for the graph. That is, the addition of variables to a display can inform and possibly change the interpretation of the display. For example, suppose one wanted to estimate the relationship between the time needed to complete an exam and the likelihood of passing. One way to estimate that relationship might be through the use of a logistic regression model. Figures 17.1a through 17.1e show a progression of graphs that illustrate different ways of displaying the results of a logistic regression analysis based on the common data set from a credentialing program used by several authors in this book, where pass rates are estimated based the amount of time candidates took to finish the exam. Figure 17.1a shows a plot of the fitted regression model for those data as a smooth, s-shaped, downward trend that is interpreted to mean that the longer candidates spend on the exam, the worse their likelihood of passing. Candidates who finish in about one hour tend to have a near 100% pass rate, whereas candidates who take more than 4 hours have pass rates of less than 50%. One interpretation of the fitted model might be that candidates who know the material well will answer questions quickly, and that candidates who do not know the material well will struggle to answer questions, taking more time. However, one might question if this trend would hold for candidates who take an unusually short time to complete the test (e.g., less than 1 hour). Figure 17.1a provides no information about how well this statistical model actually fits the observed data. Therefore, taken on its own, the graph provides an explanation of the relationship between test-taking time and pass rates, but provides no guidance regarding the accuracy of the explanation.

Figure 17.1b attempts to address this issue by overlaying the raw data on top of the fitted statistical model. Each dot represents one test taker, with candidates who passed plotted at 100% and candidates who fail plotted at 0%. This graph contains additional data, but arguably provides no additional information for the graph user. The problems are twofold. First, because of the number and specificity of observations (i.e., because the exam is computer administered, candidates' test-taking times are known to the nearest second), the data points overlap to such an extent that no conclusions can be drawn. Second, there is a mismatch in the scale between the statistical model (which provides quantitative estimates of the likelihood of passing) and the raw data (which represents a dichotomous pass/fail decision).

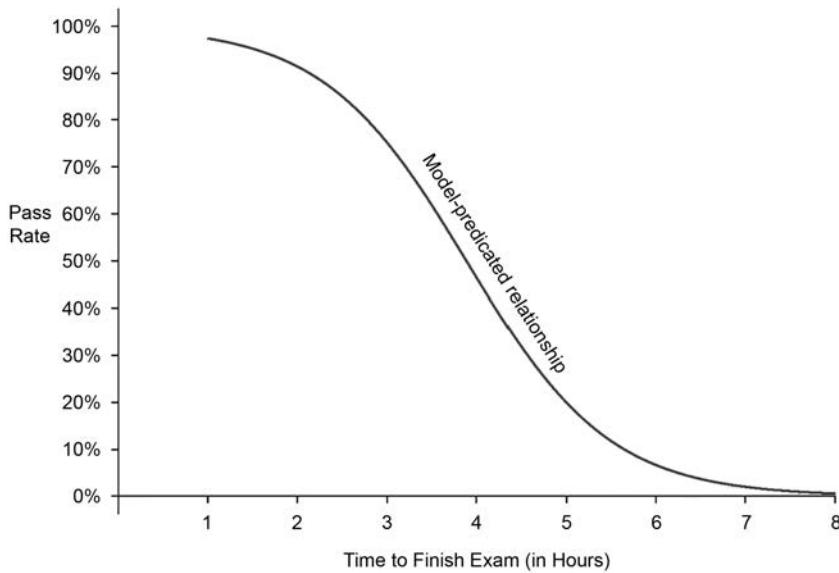


Figure 17.1a A Fitted Logistic Regression Model for Estimating Pass Rates Using Test-Taking Time; No Raw Data

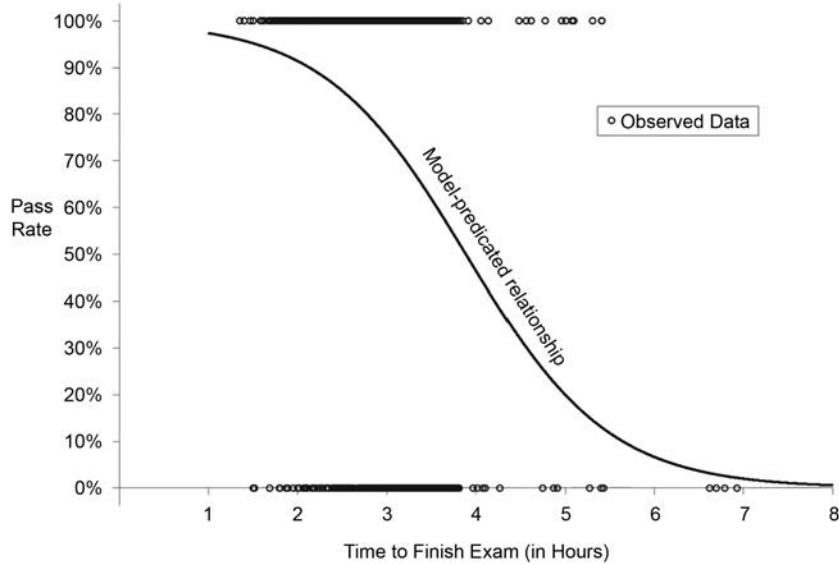


Figure 17.1b Displaying Raw Data Overlaid on a Fitted Logistic Regression Model for Estimating Pass Rates Using Test-Taking Time; Each Observation Represents a Candidate, 100% = pass, 0% = fail

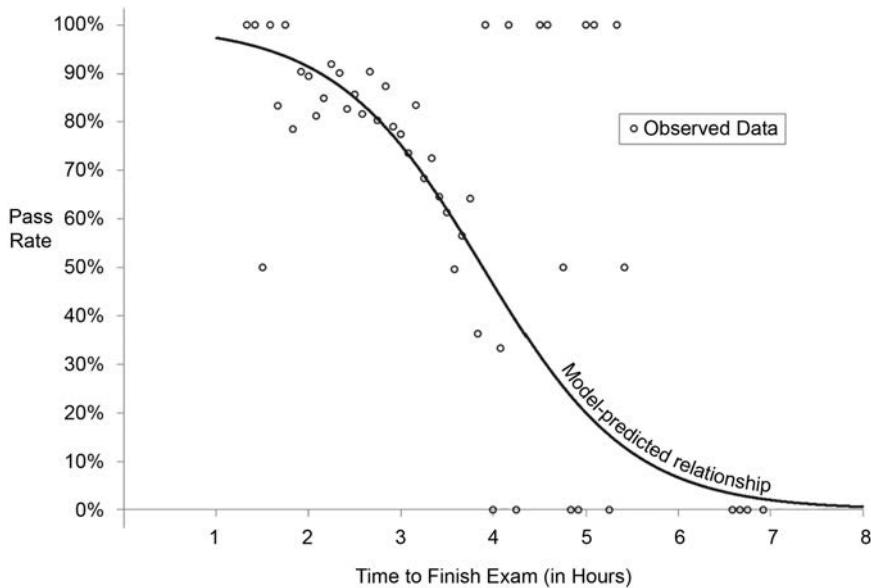


Figure 17.1c Displaying Raw Data Overlaid on a Fitted Logistic Regression Model for Estimating Pass Rates Using Test-Taking Time; Test-Taking Time Rounded to the Nearest 5 Minutes, Candidates Grouped, and Pass Rates Calculated

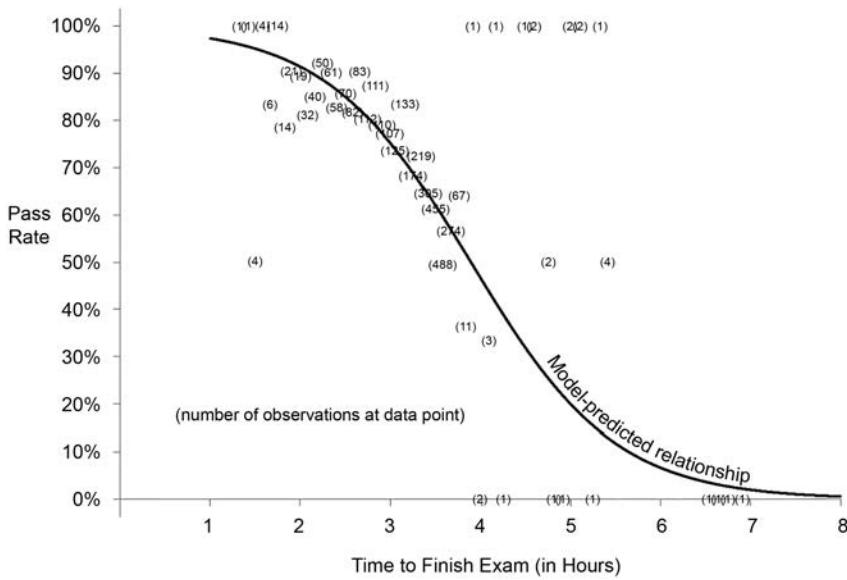


Figure 17.1d Displaying Raw Data Overlaid on a Fitted Logistic Regression Model for Estimating Pass Rates Using Test-Taking Time; Observations Replaced with the Number of Candidates at Each Data Point

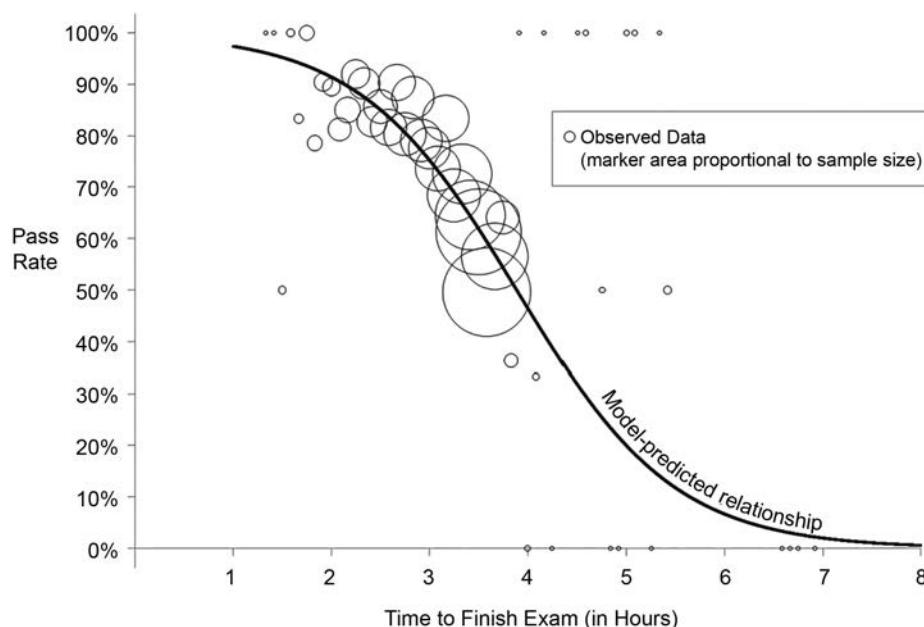


Figure 17.1e Displaying Raw Data Overlaid on a Fitted Logistic Regression Model for Estimating Pass Rates Using Test-Taking Time; Data Points Replaced by Markers with Areas That Are Proportional to the Number of Candidates at That Point

Figure 17.1c addresses the problems of Figure 17.1b by making two modifications to the raw data. First, the test-taking times are rounded to the nearest 5 minutes. Next, pass rates are calculated for each rounded time point. The result is an overlaid scatterplot that now begins to show the fit of the data to the statistical model. At first glance the model seems to do a reasonable job describing the data, but there are several points that fall far from the regression line. However, the picture shown by Figure 17.1c is incomplete and potentially misleading: The way the graph is drawn gives the impression that all data points have equal weight. The problem with this representation is that some of the points are derived from a large group of candidates (e.g., the points near the middle of the time distribution), whereas others are derived from only a single candidate.

One way to address the shortcomings of Figure 17.1c is to replace the data points with numbers indicating the number of observations comprising each data point, as in Figure 17.1d. In this way we have added an additional variable to the graphic that can help the user interpret the adequacy of the fitted statistical model. For each raw data point, the horizontal position indicates the test-taking time, the vertical position indicates the pass rate, and the plotted symbol (in this case, a number) indicates the number of candidates whose test results were used in identifying the data point. From this display, we can see that points lying near the fitted model tend to be derived from large groups of examinees; points lying far from the fitted model tend to be derived from small groups of examinees. This added information is very helpful for evaluating the fit of the statistical model. However, stylistically, the data density is sufficient to make the use of the numbering variation less than ideal. In some areas of the graph, it is difficult to read the values because they overlap with each other, the fitted model, or the graph axis.

Figure 17.1e may be a more effective solution to identifying the relative amount of information that went in to calculating each of the raw data points. This graph overlays

bubbles, with areas proportional to the number of candidates whose data contribute to each data point. This graph provides less detail than Figure 17.1d, because actual values are no longer shown. However, for many viewers of the graph, the precise values likely will be less important than generally being able to determine which points are derived from many candidates and which are derived from few. This variation of the graph, along with Figure 17.1d, shows that the model generally fits the data; however, there may be some model miss-fit for test-taking times shorter than 2 hours, indicating that the data may follow a more complex model than the one that has been fitted.

Choosing an appropriate display is an important step in visual display design. The preceding example shows how the choice of which data to display can also affect the interpretation and utility of the display. It also provides an illustration of the thought processes in which one should engage when selecting the format for a data display.

Aim for Simplicity

Visual displays can help to reduce cognitive load for users of the display (Pastor & Finney, 2013), aiding in the understanding of complex data relationships. Unnecessarily complicated or cluttered displays undermine this potential benefit. Therefore, visual displays should contain adequate detail to convey the information that the display creator intends to communicate, but should do so in a way that minimizes distracting elements. Emphasis should be placed on the data-based elements of the display, and the non-data elements should be minimized or removed (Tufte, 2001). The number of fonts, symbols, and colors used should be kept to a minimum, varied only when necessary to help with understanding. It is often desirable to produce multiple displays rather than attempting to put too much information into a single display.

For example, Figure 17.2 shows pass rates for the credentialing program data set broken down by attempt and form number. The graph is rendered in three dimensions (3-D), which may be visually appealing to some, but makes the graph unnecessarily complex and reduces readability. By making the bars 3-D, it is difficult to determine where the tops of the bars are on the graph's vertical scale. Additionally, the graph contains an excessive number of horizontal gridlines extending out from the vertical scale. If

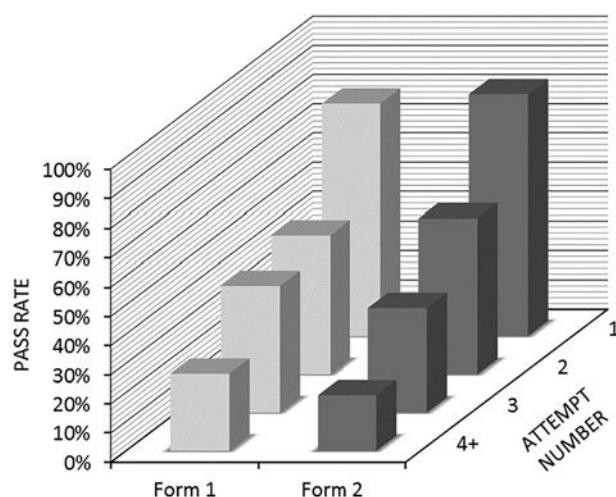


Figure 17.2 A Three-Dimensional Bar Graph Illustrating Pass Rates by Attempt Number and Test Form

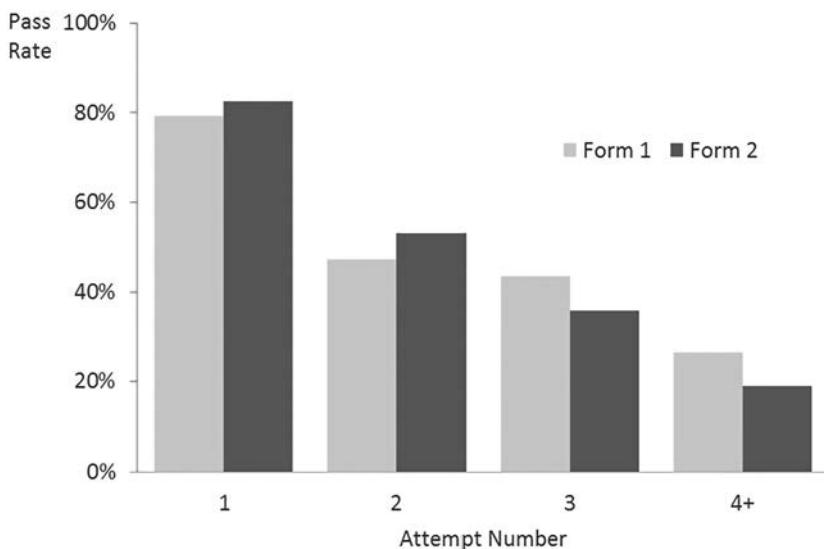


Figure 17.3 A Two-Dimensional Bar Graph Illustrating Pass Rates by Attempt Number and Test Form

the graph is intended to show broad trends in the data, this level of detail is unnecessary; if very specific values were needed, a table would be a better choice. If it is necessary to include gridlines, the lines should be thin and faint so as to be available but unobtrusive.

Figure 17.3 shows the same data in a simplified bar graph. This variation of the graph makes it easy to see the downward trend in pass rates as attempts increase. It also makes for easy and accurate comparisons across forms and attempts. The axes and axis labels are drawn in a lighter shade in Figure 17.3 than in Figure 17.2, giving less emphasis to these nondata elements, while still conveying necessary information for interpreting the graph.

The principle of simplicity applies to tables as well. That is, the use of borders, shading, and fonts should be used/varied only as necessary to aid in interpretation. A full set of heavy borders is not necessary and can make a table less readable. Table 17.1, shown previously, uses only a minimal amount of borders. Top and bottom borders delineate the beginning and end of the table. Faint horizontal guides help to separate distinct security issues, and light shading helps to place emphasis on the area(s) of primary threat.¹

Integrate Text, Numbers, and Figures

The richness of information provided by a visual display can be enhanced by combining text and numbers with graphical components of the display. The text may be part of the data, or may take the form of interpretive information or labels. Interpretive information should be proximate to the display and appropriate for the level of technical sophistication of the intended audiences. Labels and other descriptive elements also should be placed in close proximity to the feature that they describe.

For example, in Figure 17.1, the predicted statistical model is labeled directly in each panel, as opposed to having the label appear in a legend. Another simple example can be seen in Figure 17.3. The legend for the graph is included without a border inside the body of the graph itself to save space and keep the information in close proximity to the data bars. The proximate nature of the legend to the data elements keeps the

important information more accessible than if the legend had been placed outside of the plot area (e.g., to the right of the graph or below the horizontal axis label).

Table 17.2 integrates graphs and text through the use of sparklines. Sparklines are small graphs integrated into text or a table (Tufte, 2001). In this case, the sparklines are small bar graphs that allow for quick comparisons of pass rates across attempts within each country. This style of display is an efficient way to show detail alongside broad patterns. The use of a table allows precise information about the specific pass rates, while at the same time broad trends in pass rates can be easily determined for each country within a single eye span. It is clear from the table that pass rates tend to drop sharply after the first administration for candidates who went to school in the United States, and that pass rates tend to remain relatively flat across attempts for candidates who were schooled in India or the Philippines.

Table 17.3 shows a second illustration of sparklines, but in this case the visual display is summarizing data at the item level. In this display, plots of moving averages are used in concert with item difficulty measures calculated at different times to identify potential item breaches/exposures. Item difficulty values are shown as *p*-values (i.e., the proportion of candidates answering the question correctly). Small arrows are included alongside the values for *p*-value change to give a gross, at-a-glance indication of how the difficulty of the item has changed over time. The moving average plots provide a more nuanced view of the changing performance of each item.

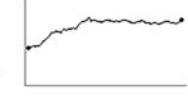
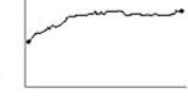
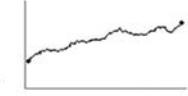
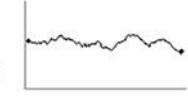
The interpretation of a visual display will not always be obvious to the reader, especially readers from nontechnical backgrounds. It is often advantageous and preferable to add interpretive information directly to the visual display itself. Figure 17.4 provides an example of the integration of numbers, text, and figures (based on a similar display from the National Center for Education Statistics, 1996, p. 35). The figure shows a comparison of pass rates for first time and repeat test takers trained in several different countries (using the common credentialing program data). Specific values are included for all pass rates. Treating the data as a sample from a larger population, differences in pass rates are presented as 95% confidence intervals, as opposed to point estimates. Because nontechnical audiences often have difficulty interpreting statistical jargon (Hambleton & Slater, 1997; Impara, Divine, Bruce, Liverman, & Gay, 1991), the display includes a set of instructions that explain the display and lead users to a correct interpretation of the data. Although the amount of text accompanying the display is more than one would expect to find for a figure in an academic journal, the placement of explanatory text

Table 17.2 An Illustration of Pass Rates by Attempt for Several Countries Using a Combination of Numbers and Sparklines

	Attempt				Trend
	1	2	3	4+	
All Candidates	81%	50%	40%	23%	
Country where candidate was educated					
USA	88%	63%	48%	25%	
India	37%	37%	38%	24%	
Philippines	37%	36%	28%	17%	
Other	42%	25%	27%	27%	

Note: Highest pass rate for each country is shown in **bold**.

Table 17.3 Item Performance for a Subset of Items Across the First 200 Administrations of Form 1

Item	<i>p</i> -value			
	After 50 admins	After 200 admins	Change	Moving Average
7	0.42	0.80	0.38	
12	0.42	0.36	-0.06	
19	0.52	0.92	0.40	
38	0.24	0.70	0.46	
60	0.30	0.80	0.50	
158	0.56	0.42	-0.14	

Note: This table integrates graphs with numbers through the addition of sparklines to show how item *p*-values change with the number of test administrations. *P*-value refers to the proportion of candidates who answered the question correctly. "After 50 admins" is the *p*-value calculated for the first 50 test administrations. "After 200 admins" is the *p*-value calculated for administrations 151 through 200. The moving average plot shows the average across the first 200 administrations with a period of 50 administrations.

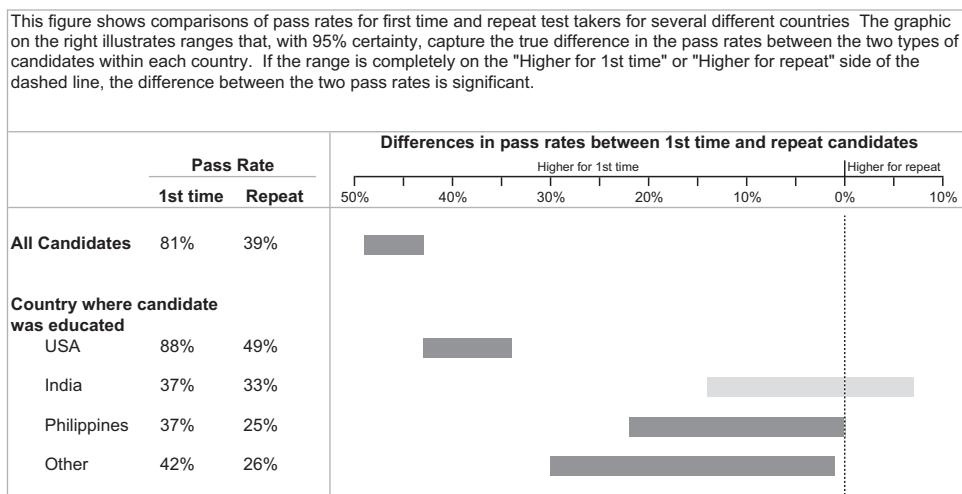


Figure 17.4 Comparison of Pass Rates for First Time and Repeat Test Takers, for several Countries (design based on National Center for Education Statistics, 1996, p. 35)

adjacent to the display (as opposed to a separate interpretive guide) increases the likelihood of the audience understanding and interpreting the display correctly.

Highlight What Is Important

The guideline to highlight important information is complementary to the two previous guidelines (i.e., aim for simplicity; integrate text, numbers, and graphs). The reader's eye should be drawn to the salient features of the visual display. This can be partially accomplished through the minimization of nondata elements and enhanced by placing emphasis on specific, important features. Design features such as color, shading, labeling, and special symbols also can be used alone or in combination to add emphasis.

Table 17.2 highlighted important observations through the selective use of bold text and shading. For each country, the highest pass rate is highlighted, and the corresponding bar in the associated bar graph is rendered in a darker shade of gray. This emphasis makes it easy to see that India is unusual compared to the other countries in that the highest pass rates are for the third attempt, as opposed to all other countries where pass rates are highest for the first attempt.

Figure 17.5 shows a scatterplot comparing test-taking time and test scores for one form of the common credentialing program data. Although this graph does a reasonable job showing the shape of the pattern of the relationship between the variables, there is no emphasis given to any one area of the data. If, however, a testing program wished to emphasize test administrations where candidates received very high scores in very short periods of time (say, flagging administrations with scores at or above the 95th percentile, and test-taking times at or below the fifth percentile), Figure 17.6 would be a better choice. This graph has several features that immediately draw the readers' attention to the flagged administrations: First, guides are added to indicate the relevant percentiles. The guides also serve as a means for dividing the display into quadrants. Next, marker styles are changed to added emphasis. Flagged observations are given their own marker style (heavy plus signs), and the nonflagged observations (circles) are rendered in a lighter shade to reduce emphasis. Finally, an interpretive label is placed

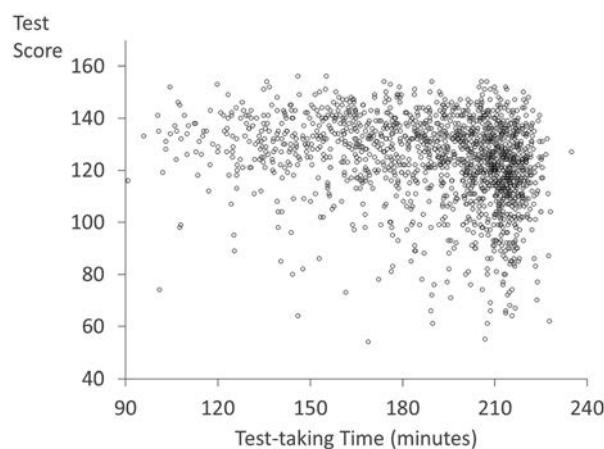


Figure 17.5 A Scatterplot of Test-Taking Time (in minutes) Versus Test Score for Form 1, Showing the Overall Pattern for the Relationship Between the Two Variables

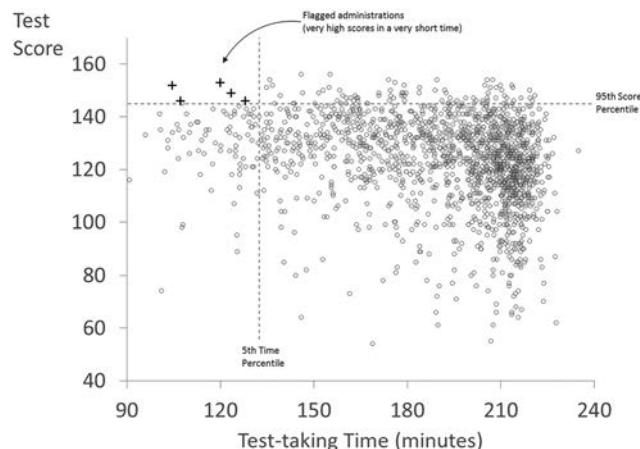


Figure 17.6 An Annotated Scatterplot of Test-Taking Time (in minutes) Versus Test Score for Form 1, Emphasizing the Flagged Administrations

directly in the body of the graph. The label “Flagged administrations (very high scores in a very short time)” adds additional emphasis and provides interpretive information for the reader, avoiding the necessity of a legend or for the reader to find the interpretation elsewhere in the text. The label identifies the observations as flagged and provides information as to why they were flagged.

Several of the figures use shades of gray to differentiate between groups (e.g., Figures 17.2 and 17.3) and to add emphasis. Color can also be an effective way to achieve the same goals. When using color in displays, large areas of data should be produced using a muted color palate. Vivid/saturated colors should be reserved for areas of emphasis. Colors should be distinct enough from one another to be clearly differentiable, and care should be taken to ensure that the distinctions are still clear if printed in black and white (if it is likely the display will be printed). Albers (2013), Few (2012, pp. 77–79, 344), Tufte (1990, pp. 81–95), and Wong (2013, pp. 40–47) each provide additional information and guidance for including color in visual displays.

Do Not Intentionally Mislead

This guideline may seem obvious, however, it is especially important in the context of test security investigations. Actions taken against candidates or groups suspected of fraud can have serious or wide-ranging consequences. When quantitative evidence is brought before a review panel, arbiter, or judge, it is important to avoid the appearance of misrepresentation. Visual displays should be constructed to avoid distortions and ambiguities. Because it is seldom the case where one can say with 100% certainty that cheating has taken place, professional judgment needs to be brought to bear in most situations. The individuals making those judgments need clear, unbiased information to make informed decisions. It was noted earlier that creators of visual displays should aim for simplicity, removing or minimizing nonessential elements from the display. However, care should also be taken to avoid removal of visual elements that are necessary for the correct and accurate interpretation of the display.

One of the ways graphs can be used to manipulate is through the inappropriate choice of vertical and/or horizontal scales. Few (2009, p. 93) provided a set of rules of thumb that can help to avoid this problem: (1) the scale of bar graphs should always

be based at zero.; (2) for other types of graphs, the scales should begin at a point just below the lowest value and extend just above the highest value; and (3) the end points of the scale should be round numbers and intervals along the scale should be round numbers that make sense in the context of the scale. Few also notes that when a scale does not begin at zero (for a graph other than a bar graph) this fact should be highlighted for the reader (e.g., with a footnote) if there is concern that the reader might be misled by the scale (2012, p. 193). The following example provides an illustration of how scale choice can affect the interpretation of a graph.

Arguably one of the highest-stakes cases related to a cheating investigation is that of *Vela v. Nebraska* (2006). In this case, the death penalty was being considered for a defendant who had been convicted of murder during a bank robbery. The defense was attempting to make the case that the defendant was ineligible for the death penalty under *Atkins v. Virginia* (2002), in which the U.S. Supreme Court ruled that individuals with mental retardation are not eligible for the death penalty. At the time of the trial, there was a Nebraska statue indicating that there would be a presumption of mental retardation for individuals with an IQ of less than 70. The defendant had taken several IQ tests, some of which indicated a score of greater than 70 (early in the case timeline, when it was not clear if the death penalty would be an option) and some with a score less than 70 (after it became clear that the defendant likely would be eligible for the death penalty). The judge requested the assistance of a psychometrician to help reconcile the IQ score differences (see Buckendahl & Foley, 2011, for a more detailed description of the case). An important cheating-related issue from this case was whether or not the defendant had maledgered (i.e., intentionally performed worse than he was able) on one or more of the IQ assessments.

One piece of relevant quantitative information from this example is the defendant's IQ scores on the three assessments. Figure 17.7 shows a line graph of the three test scores. This graph is designed to emphasize the difference between test scores. The vertical scale, centered closely around the data, the narrow format (taller than it is wide) of the graph, and the equally spaced observations across the bottom of the graph make it appear to a casual observer that there was a precipitous drop in the defendant's test scores, going from very high to very low. One problem with this graph is that the time between the assessments is ignored. Figure 17.8 improves the graph by adding a time scale to the horizontal axis. In addition to the time scale, the graph is widened slightly. As a result, the pattern of the defendant's scores appears to decrease less quickly. Figure 17.8 meets Few's (2009) rules of thumb for scale selection, and it could be argued that this graph is a fair depiction of the IQ scores over time.

However, one may still question the vertical scale of Figure 17.8. The high value for the scale is 90 and the low value is 65. These values may be misinterpreted by individuals not familiar with IQ scores. That is, IQ scores are scaled in such a way that a score of 100 represents average intelligence, and the score scale has a standard deviation of 15 points. Therefore, all scores in the 65–90 point range shown in Figure 17.8 are below average. Figure 17.9 presents an alternative format for the graph. In Figure 17.9, the vertical scale ranges from 55 to 145; this centers the scale at the average score and extends the scale to three standard deviations above and below the mean (a range which would encompass the IQ scores of the vast majority of the U.S. population). The vertical scale labels are set in 15-point (i.e., one standard deviation) increments. The plot area is also stretched horizontally into a somewhat more conventional rectangular shape. By extending the vertical scale and changing the shape of the plot area, the test score data is relegated to only the lower half of the graph, and the changes in scores over time appear less substantial. Depending on one's point of view, these changes may appear inappropriately

manipulative. However, one could argue that the extension of the vertical axis is useful in that it does a better job of placing the defendant's scores in context. The revised vertical scale, along with the addition of labeled lines indicating "average intelligence" and the statutory value indicating "mental retardation," gives a more holistic version of the defendant's scores: all of his scores were below average, but early scores indicated that he was above the statutory value for mental retardation, whereas for the last assessment, his score fell below the threshold. This version of the graph may be more appropriate for audiences who are less familiar with the IQ score scale.

Figures 17.7, 17.8, and 17.9 are all based on the same data. Clearly, the choice of scales can have a dramatic effect on the appearance of a visual display and may change the way the display is interpreted. This example is not meant to identify a "correct" scale choice, only to highlight the influence that the scale choice can have on the interpretation. Creators of visual displays should think critically about scale and other design choices that have the potential to intentionally or unintentionally mislead the display user. Care should be exercised to ensure that all design choices are defensible. The best choice of scale for the above example may depend on the level of familiarity the intended audience has with the IQ score scale, which leads to the next design principle: pair the design with the audience.

Pair the Design With the Audience

Several researchers (e.g., Carswell & Ramzy, 1997; Impara et al., 1991; Zwick, Zapata-Rivera, & Hegarty, 2014) have found that understanding of and preferences for different types of visual displays varies depending on the characteristics of the audience (e.g., background knowledge, gender, education). Therefore, audience should be a primary consideration when creating visual displays. Audiences without strong technical backgrounds may have trouble understanding statistical concepts, such as confidence

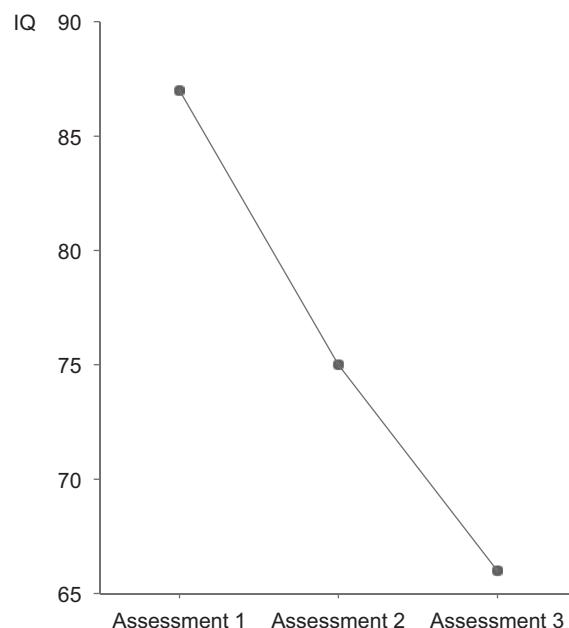


Figure 17.7 Line Graph Depicting Three Assessment Scores

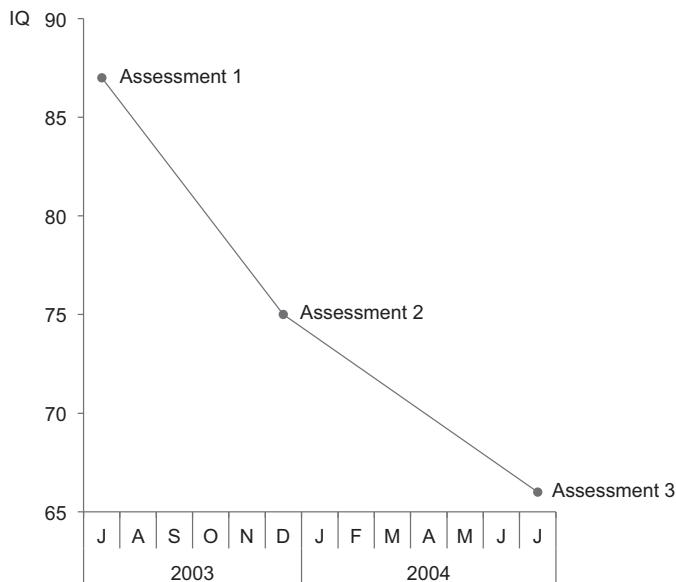


Figure 17.8 Time Series Graph Depicting Three Assessment Scores Measured over Time

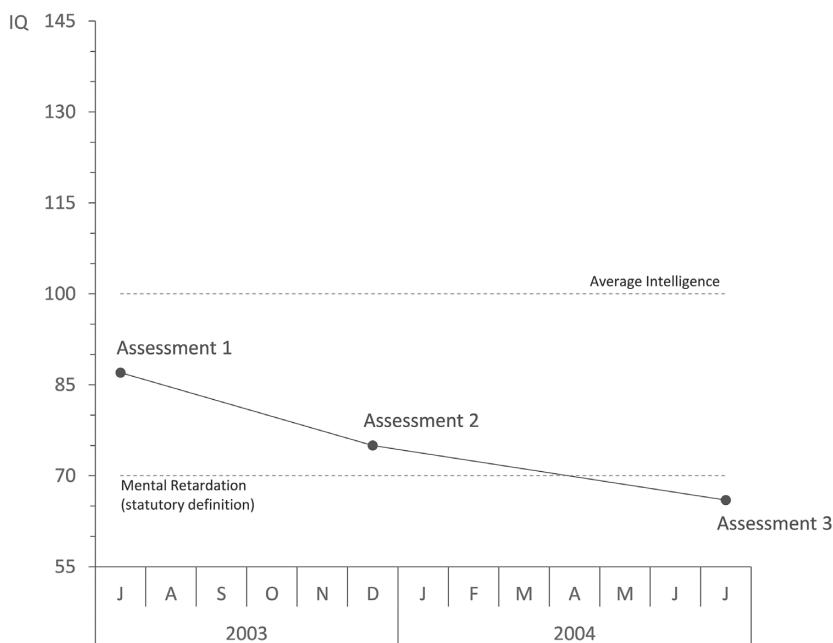


Figure 17.9 Time Series Graph Depicting Three Assessment Scores Measured over Time, with Rescaled Axis and Important Scale Values Labeled

intervals. In cases where these more complex concepts are used, it may be necessary to include additional interpretive information within the display (see Figure 17.4). As seen in the previous example, Figure 17.8 may be perfectly acceptable if the audience has a strong understanding of the IQ score scale, whereas Figure 17.9 may be more appropriate for audiences with less background knowledge. If the background knowledge of

the intended audience is mixed or unknown, it may be necessary to include variations of a display and/or err on the side of imbedding additional interpretive information.

Summary

The examples provided in the previous sections illustrate the design principles discussed, but can also serve as prototypes for other quantitative fraud detection metrics. For example, Table 17.3 could be modified to show comparisons of form or test center statistics (e.g., pass rates, mean scores) over time (as opposed to items) to help illustrate potential security breaches. Scatterplots with identified thresholds, such as Figure 17.6, can be modified to display other measures, such as person-fit metrics or changes in test scores (e.g., initial vs. retake scores with each observation representing a candidate).

The previous sections identified six broad themes synthesized from the visual display literature. These themes embody design principles that can be applied to a variety of quantitative methods for fraud detection. However, there are other aspects of visual display design that are especially relevant to assessment results in general, and fraud detection in particular: inclusion of indicators of uncertainty and conveying information about probabilities. The following two sections discuss these important issues.

INDICATORS OF UNCERTAINTY

Whenever metrics are calculated based on sample (as opposed to the full population of interest), the resulting values contain sampling error. In this context, *error* is not referring to a mistake, but rather to a degree of uncertainty. For example, when we calculate a pass rate for a testing site using data from a 2-week testing window, data from this window can be thought of as a sample. If we had collected data during a different time frame, the estimated pass rate would likely be similar, but not identical, to the value obtained from the initial sample. Sampling error is a quantification of this sample-to-sample variability.

In the same way, the performance of a candidate on an exam (e.g., the candidate's raw or scale score, performance on various subscales) is also based on a sample. That is, the items or tasks on the exam are based on a sample from the population of items/tasks that could be used to measure the candidate's competence in whatever content domain that the exam is measuring. For example, the 170 items on Form 1 of the common credentialing program data set, though selected based on a set of test specifications, represent a sample from the population of possible items that could have appeared on the exam. If a candidate took a different set of 170 items, built to the same specifications, the candidate would likely receive a similar, but not identical, score. This variability in performance is central to the psychometric concept of measurement error.

Sampling and measurement error both introduce uncertainty into test results. If visual displays are created that do not address these sources of error, users of these displays may overestimate the level of precision (i.e., underestimate the uncertainty) inherent in the data. Therefore, if indicators of uncertainty are omitted from visual displays, the display may be inadvertently misleading users. Wainer (2009, p. 121) recommends that effective visual displays should remind users that the data displayed contain uncertainty, quantify the magnitude of the uncertainty, and help users avoid misinterpretations based on incomplete understandings of the uncertainty in the data.

Sampling error will often arise in fraud detection when presenting results for one or more groups of candidates. Figure 17.10 shows a bar graph displaying the average

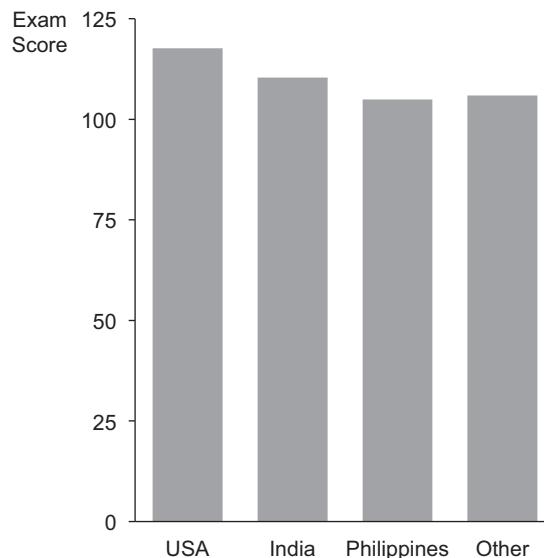


Figure 17.10 Mean Exam Score for Form 1 Repeat Test Takers, by Country

scores of repeat test takers for Form 1 of the common credentialing program data set for candidates who were trained in several different countries; the graph contains no indicators of uncertainty. If the data from the example credentialing program is thought of as a sample of the larger population of candidates, Figure 17.11 might be a better choice for displaying mean scores. In Figure 17.11, mean scores are displayed on the graph as points accompanied by error bars representing two standard errors (of the mean) above and below each data point. This representation helps to avoid misinterpretations: for example, there is substantial overlap in the error bands for retakers from India and the Philippines, indicating the difference may not be statistically significant (i.e., larger than one would expect by chance). Lane and Sándor (2009) advocate for use of graphs like that shown in Figure 17.12. This graph shows box-and-whisker plots for each country. While forgoing the presentation of specific information about sampling error, these plots give a wealth of distributional information (e.g., medians, quartiles, ranges, outliers) that is not included in Figures 17.10 and 17.11. Sampling error was also shown in Figure 17.4. Whereas Figure 17.11 shows error bands about the means for each group, Figure 17.4 shows error bands about the difference between means for two groups (in this case first time and repeat test takers within a given country).

Although sampling error is often associated with statistics for groups of candidates, measurement error can be especially important to evaluate when considering scores of individuals. Recall the death penalty example discussed above. A primary issue noted in this example was the differences in the defendant's scores across three different intelligence tests. Each of these tests represented a sample of items and/or tasks selected from a larger population of potential items/tasks; therefore, each test score is subject to measurement error. Figure 17.13 expands on Figure 17.9 through the addition of error bars that indicate measurement error by showing two standard errors of measurement above and below each observed test score. This version of the graph gives a more complete picture of the change in performance across the three assessments, and, through the addition of error bars, gives additional information about the relationships between the scores and the statutory definition of mental retardation.

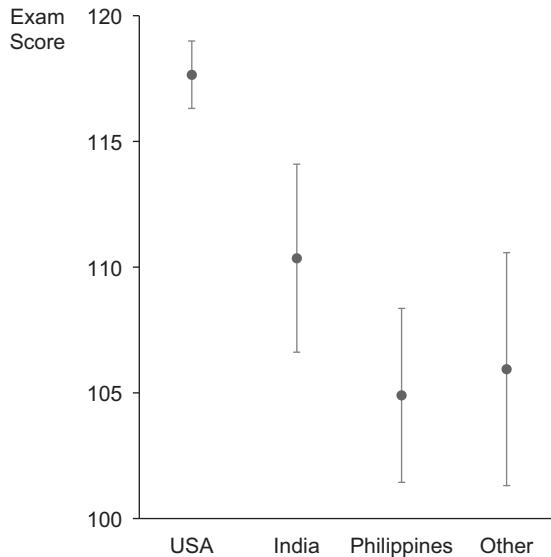


Figure 17.11 Mean Exam Score for Form 1 Repeat Test Takers, by Country. Error Bars Indicate Two Standard Errors (of the mean) Above and Below the Mean

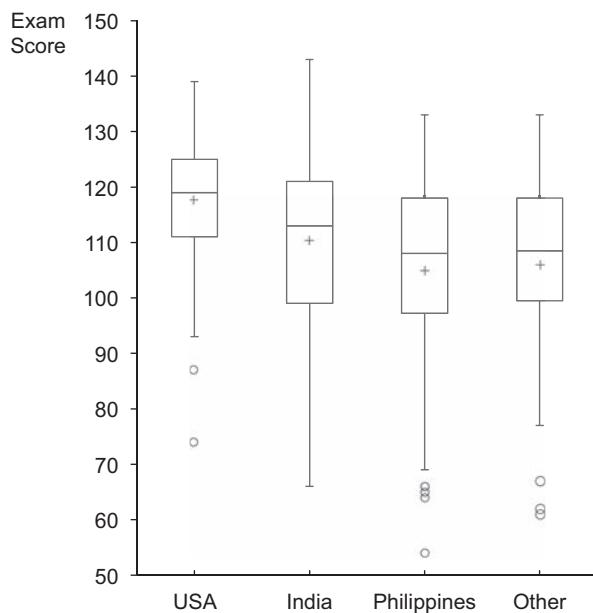


Figure 17.12 Exam Score Distribution for Form 1 Repeat Test Takers, by Country

Including indicators of uncertainty in visual displays sometimes brings other design principles into conflict: in attempting to avoid misleading display users, the display may become too complex for nontechnical users to understand. That is, error bars and other indicators of uncertainty reflect sophisticated statistical concepts that may be unfamiliar to audiences without a strong technical background. Therefore, when such

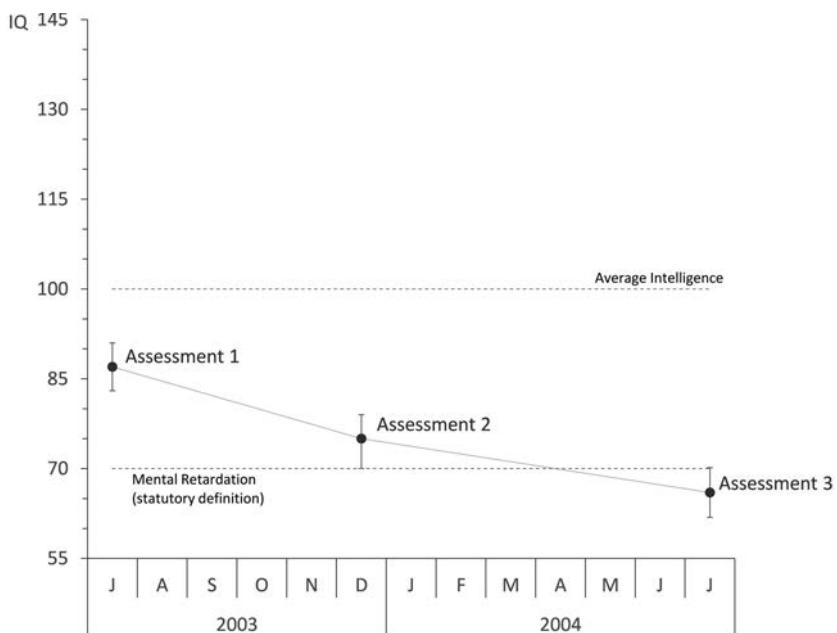


Figure 17.13 Time Series Graph Depicting Three Assessment Scores Measured Over Time, With Important Scale Values Noted and Error Bars That Indicate Two Standard Errors of Measurement Above and Below the Observed Test Score

indicators are included, special care should be taken to ensure that sufficient interpretive information (e.g., labels, embedded description) is included for examinees to correctly interpret the displays. Additional information about ways to depict error associated with assessment results can be found in Foley (2015) and Wainer (1996, 2009, Ch. 13). See Zwick et al. (2014) for an example of an empirical study comparing different ways to visually and verbally depict measurement error.

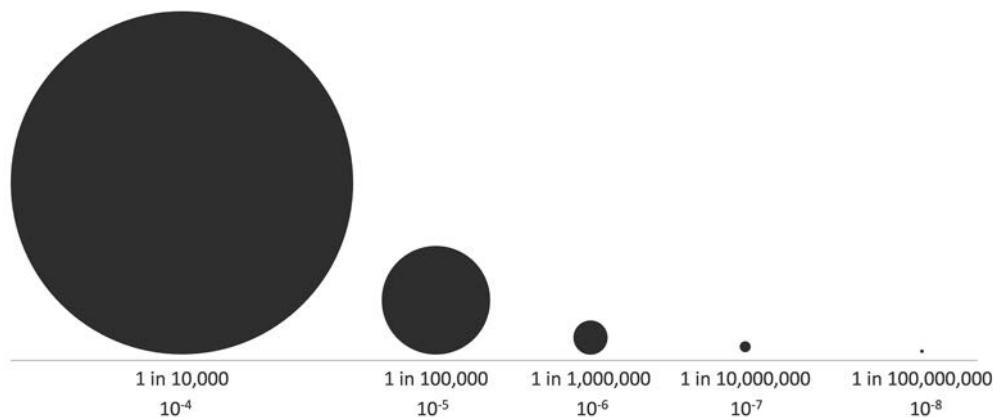
DEPICTING PROBABILITIES

Many methods of statistical detection of test fraud rely on probabilistic determinations of aberrant performance. It can be difficult for nontechnical test users to understand probabilistic reasoning. (For example, what does it mean when we observe two test takers with a level of agreement that would only have a probability of 3×10^{-7} of occurring by chance?) Specifically, when test fraud is detected, the metrics involved often include infinitesimal values (indicating events extremely unlikely to occur due to chance) that most users rarely encounter in everyday life; consequently they have little context for interpreting such values. One way to assist users in understanding these probabilities is by providing multiple, equivalent explanations of the probabilities in question. For example, Table 17.4 presents several different formats for presenting probabilities. By including such a table when discussing fraud metrics that deal with probabilities, the likelihood of user understanding is increased by giving them a translational tool for converting probabilities in unfamiliar formants into terms that they may find more familiar.

Because extremely large and extremely small numbers often fall outside of users' everyday experiences, it may be difficult for them to recognize the immense differences that exist between observations separated by several orders of magnitude. For example, nontechnical users may not understand the substantial difference between 10^{-4} and

Table 17.4 Alternative Formats for Expressing Probabilities

Value	Written	Decimal Notation	Scientific Notation
1 in 10	one in ten	0.1	10^{-1}
1 in 100	one in one hundred	0.01	10^{-2}
1 in 1,000	one in one thousand	0.001	10^{-3}
1 in 10,000	one in ten thousand	0.0001	10^{-4}
1 in 100,000	one in one hundred thousand	0.00001	10^{-5}
1 in 1,000,000	one in one million	0.000001	10^{-6}
1 in 10,000,000	one in ten million	0.0000001	10^{-7}
1 in 100,000,000	one in one hundred million	0.00000001	10^{-8}
1 in 1,000,000,000	one in one billion	0.000000001	10^{-9}

**Figure 17.14** An Illustration of the Relative Magnitude of a Select Set of Probabilities. The Areas of the Circles Represent the Relative Size of Each Value in Relation to the Others

10^{-8} . The values differ by only 4 units in the exponent, but in real terms, one is 10,000 times greater than the other. Illustrations like that shown in Figure 17.14 can help to make these relationships more concrete.

Similarly, when providing probabilistic fraud metrics, it can be advantageous to use visual displays to help give the values context. For example, Figures 17.15 and 17.16 illustrate the probability of two examinees who both scored 70% correct on the examination represented by the common credentialing program dataset having 27 incorrect responses in common (assuming four-option multiple choice items). The curve in Figure 17.15 shows the probability for each number of possible agreements ranging from 1 to 30, based on random chance. The curve makes it clear that the most likely number of chance agreements would be approximately five, with the typical number of chance agreements falling in the 0–13 range. The observed number of agreements (27) is indicated with an arrow. The extremely small likelihood of this number of agreements is shown graphically and is also explained in an embedded label to help users interpret the graph. Figure 17.16 is a variation of the same graph. (In this version, the cumulative probability is displayed, as opposed to the discrete probability.) For example, there is an approximately 68% chance of five or more incorrect agreements, but only a 5% chance of 10 or more incorrect agreements.

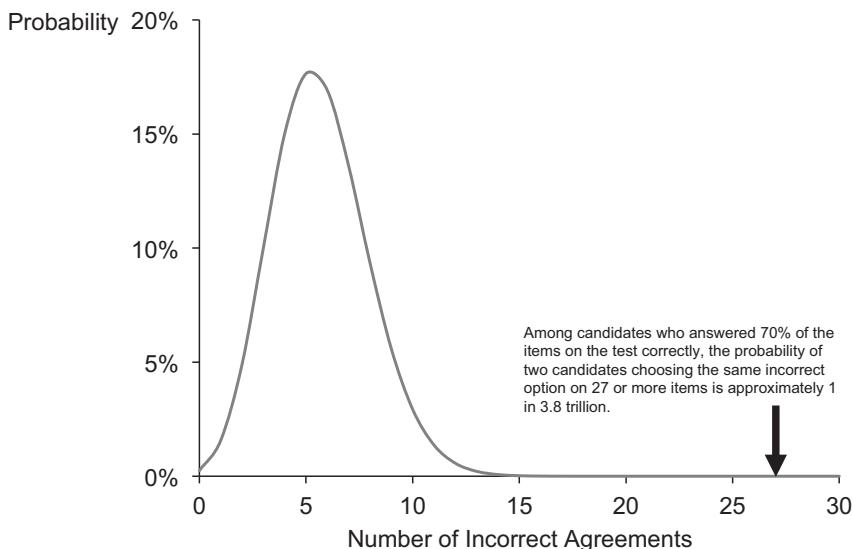


Figure 17.15 The Probability Distribution of the Expected Number of Incorrect Agreements Based on Random Chance

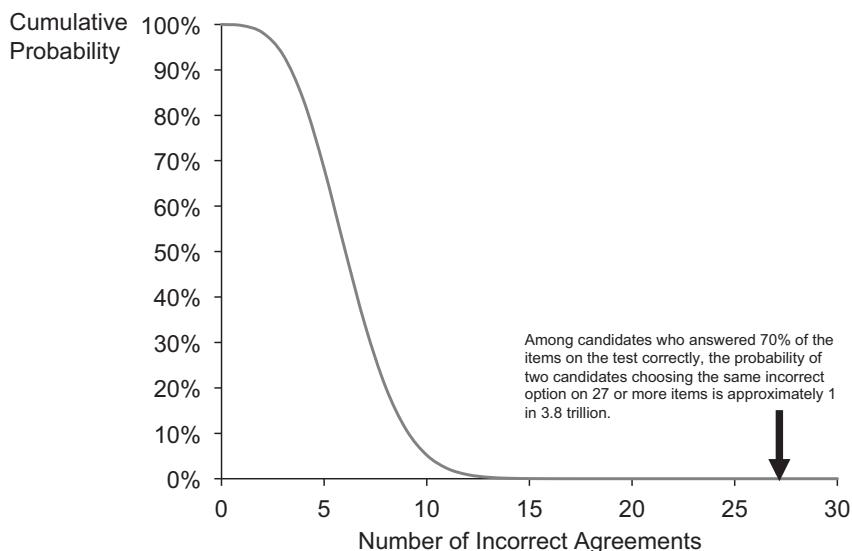


Figure 17.16 The Cumulative Probability Distribution of the Expected Number of Incorrect Agreements Based on Random Chance

Maynes (2014, p. 64) provided an example of a bivariate method for presenting probabilities for assessment data. His display was similar to a topographic map in that he created contour lines for various probabilities and overlaid observed data with a model-predicted value. A variation of his graph is shown in Figure 17.17. This hypothetical example examines the number of correct and incorrect agreements for a pair of test takers. The location of the observed number of agreements with respect to the probability contour lines indicate that the number of agreements is extremely unlikely due to chance. This may be indicative of collusion among the candidates.

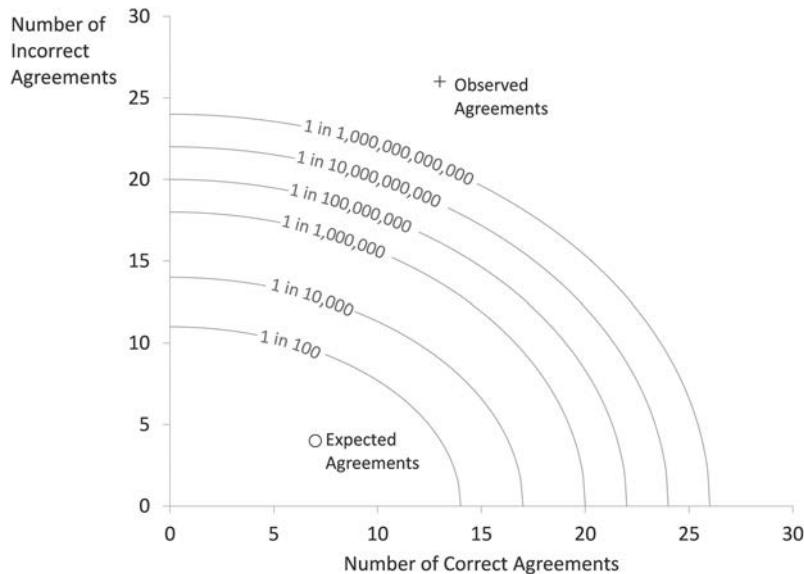


Figure 17.17 A Contour Plot Indicating the Probability of the Number of Observed Agreements Between a Pair of Hypothetical Candidates

CONCLUSIONS

Quantitative measures of test fraud are often based on sophisticated statistical techniques that can be difficult for nontechnical audiences to understand. Well-constructed visual displays can reduce cognitive load and make complicated information more accessible. This chapter outlined a set of research-based design principles that test-security professionals can apply when developing visual displays for security analyses. When using visual displays, test-security professionals should be cognizant of their audiences. The level of technical expertise held by policy makers and other stakeholders will vary, and visual displays should be adapted to the level of each intended audience to help ensure accurate interpretations of test security data.

This chapter is limited in its scope in that it focuses on broad themes related to the creation of visual displays. There is a rich literature on the art and science of visual displays, and the references mentioned in this chapter provide a wealth of information for readers who would like to explore the ideas presented here in greater depth.

ACKNOWLEDGMENTS

I would like to thank Jill Burroughs for sharing input and ideas in the early stages of this project. I would also like to thank Susan Davis-Becker and the volume editors for their constructive comments and suggestions on early versions of the chapter. Finally, I'd like to thank Jennifer Paine for sharing her unparalleled skill in proofreading. This chapter is far better thanks to the help of these talented folks. Any remaining errors or omissions are my own.

NOTE

1. It should be noted that in some cases table formats are mandated by publication style guides (see, for example, Nicol & Pexman, 2010).

REFERENCES

- Adams, B. (2014, March 1). Security: It's about validity! [Web log post]. Retrieved from <http://blog.alpinetesting.com/security-its-about-validity/>
- Albers, J. (2013). *Interaction of color* (4th ed.). New Haven, CT: Yale University Press.
- Atkins v. Virginia*, 536 U.S. 304 (2002).
- Bertin, J. (1983). *Semiology of graphs* (W. Berg, Trans.; H. Wainer, Technical Ed.). Madison, WI: University of Wisconsin Press. (Original work published 1973.)
- Buckendahl, C. W., & Foley, B. P. (2011). High stakes uses of intelligence testing. In J. Bovaird, K. Geisinger, & C. Buckendahl (eds.), *High stakes testing in education: Science and practice in K-12 settings* (pp. 191–210). Washington, DC: American Psychological Association.
- Carswell, C. M., & Ramzy, C. (1997). Graphing small data sets: Should we bother? *Behaviour & Information Technology*, 16(2), 61–71.
- Cleveland, W. S. (1985). *The elements of graphing data*. Monterey, CA: Wadsworth.
- Few, S. (2009). *Now you see it: Simple visualization techniques for quantitative analysis*. Oakland, CA: Analytics Press.
- Few, S. (2012). *Show me the numbers: Designing tables and graphs to enlighten* (2nd ed.). Burlingame, CA: Analytics Press.
- Foley, B. P. (2015). Tailoring visual displays to improve test score interpretation: Including indicators of uncertainty. In M. McCrudden, G. Schraw, & C. Buckendahl (eds.), *Use of visual displays in research and testing: Coding, interpreting, and reporting data* (pp. 265–298). Charlotte, NC: Information Age.
- Gillan, D. J., Wickens, C. D., Hollands, J. G., & Carswell, C. M. (1998). Guidelines for presenting quantitative data in HFES publications. *Human Factors*, 40(1), 28–41.
- Hambleton, R. K., & Slater, S. C. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). Los Angeles, CA: Center for the Study of Evaluation.
- Harris, R. L. (1999). *Information graphics: A comprehensive illustrated reference*. New York, NY: Oxford University Press.
- Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers? *Educational Measurement: Issues and Practice*, 10(4), 16–18.
- Kingston, N. M., & Clark, A. K. (Eds.). (2014). *Test fraud: Statistical detection and methodology*. New York, NY: Routledge.
- Lane, D. M., & Sándor, A. (2009). Designing better graphs by including distributional information and integrating words, numbers, and images. *Psychological Methods*, 14(3), pp. 239–257.
- Mayer, R. E. (2013). Fostering learning with visual displays. In G. Schraw, M. McCrudden, & D. Robinson (eds.), *Learning through visual displays* (pp. 47–73). Charlotte, NC: Information Age.
- Maynes, D. (2014). Detection of non-independent test taking by similarity analysis. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 53–80). New York, NY: Routledge.
- National Center for Education Statistics. (1996). *NAEP 1994 reading report card for the nation and states*. Washington, DC: Institute of Education Sciences, U.S. Department of Education.
- Nicol, A. A. M., & Pexman, P. M. (2010). *Presenting your findings: A practical guide for creating tables*. Washington, DC: American Psychological Association.
- Pastor, D. A., & Finney, S. J. (2013). Using visual displays to enhance understanding of quantitative research. In G. Schraw, M. McCrudden, & D. Robinson (Eds.), *Learning through visual displays* (pp. 387–415). Charlotte, NC: Information Age.
- Tufte, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.
- Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.
- Vela v. Nebraska*, No. CR02-236. Madison, NE: Dist. Ct. of Madison County, (2006).
- Wainer, H. (1996). Depicting error. *The American Statistician*, 50(2), 101–111.
- Wainer, H. (1997). *Visual revelations*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Wainer, H. (2005). *Graphical discovery: A trout in the milk and other visual adventures*. Princeton, NJ: Princeton University Press.
- Wainer, H. (2009). *Picturing the uncertain world*. Princeton, NJ: Princeton University Press.
- Wollack, J. A., & Fremer, J. J. (eds.). (2013). *Handbook of test security*. New York, NY: Routledge.
- Wong, D. M. (2013). *The Wall Street Journal guide to information graphics: The dos and don'ts of presenting data, facts, and figures*. New York, NY: Norton.
- Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, 19, 116–138.

18

THE CASE FOR BAYESIAN METHODS WHEN INVESTIGATING TEST FRAUD

William P. Skorupski and Howard Wainer

“I would never die for my beliefs because I might be wrong.”

Bertrand Russell

INTRODUCTION

Examinee cheating is always a concern for testing programs with high stakes (e.g., Cizek, 1999; Thiessen, 2007). There are obvious issues with fairness involved (students receiving higher scores than they deserve, contaminating criterion- and norm-referenced inferences), as well as concerns for the psychometric integrity of scores from assessments where cheating has occurred. As such, cheating is a validity issue that affects not only individual ability estimates by potentially biasing them upwards but may in fact result in other fairly earned ability estimates being biased downwards (in terms of their relative position in the score distribution).

Cheating behavior may occur in multiple ways. Many research studies over the years have conceptualized cheating in terms of collusion among examinees, aberrant response patterns characterized by lack of person fit, unexpected score gains, and suspicious answer changes or erasures (e.g., Angoff, 1974; Cannell, 1989; Meijer & Sijtsma, 1995; Wollack, 1997, 2003). Ultimately, cheating may be manifested in a number of different ways, so it is reasonable to consider multiple approaches to detection (while keeping in mind the possible inflation of Type I error that may occur).

The purpose of this chapter is not to focus on individual cheating detection methods, but rather to provide logical and analytic evidence to encourage Bayesian reasoning for cheating detection, regardless of the method employed. Many methods currently rely on traditional null hypothesis testing to flag potential cheaters for further review. We present an argument here that these frequentist inferences are often misleading (e.g., Gelman, 2011), and that this is especially true with regard to establishing the probability of very-low-incidence events (Savage & Wainer, 2008), such as cheating. Our goal is push social scientists to consider that the traditional frequentist p -value is really the wrong P . We posit that most cheating detecting investigations are not really interested

in the probability of observing data, given the null “not a cheater” condition (Gelman, 2013); rather, the inference of interest is the probability that a condition (“is a cheater”) has been met, given what has been observed in the data. This is precisely what the Bayesian paradigm provides. In practical terms, we will demonstrate in the following sections of this chapter that it is much more useful to know, for example, that there is a 90% chance that someone is a cheater, given the data, as opposed to saying that there is an infinitesimal chance of observing this person’s data if he or she were a noncheater. Statements like the latter are confusing at best, misleading at worst. In short, we advocate for using Bayesian posterior inferences instead of these troublesome *p*-values.

For any high-stakes operational testing program, it is of vital importance to incorporate powerful methods to detect cheating behavior. Consequently, cheating detection methods and their evaluation have become increasingly prevalent in the literature. However, although statistical power is of great importance, so is maintaining an acceptable Type I error rate that avoids false positives. Whereas the need to catch cheaters is great, the need to protect noncheaters from false accusations is perhaps equally important, if not more so.

This begs the question: What are the consequences associated with false positives and false negatives? That depends, in large part, on the unit of analysis and the course of action taken as a result of being identified as a possible cheater. For example, when a teacher or administrator is accused of cheating (as events in the Atlanta Public Schools case have demonstrated), the cost of being flagged as a cheater may be the loss of one’s career and criminal charges. If those are false positives (we are not suggesting they are), those would be dire consequences, indeed. The consequences for an individual examinee accused of cheating might result in having to retake a test, which doesn’t sound too harsh, or possibly being barred from further testing, which could mean exclusion from one’s chosen career. It should be obvious that extreme caution must be exercised when flagging examinees as potential cheaters. The cost of a false negative—the cheater gets away with it—may or may not be just as serious. One only has to imagine being on the operating table of a surgeon who cheated on a medical board examination to appreciate the potential seriousness of false negatives. Thus, when discussing the performance of a cheating detection statistic, one must sensibly evaluate its sensitivity and specificity in terms of how cheating behavior is flagged, and the relative costs and benefits associated with being wrong or being right.

Consider a motivating example, adapted from a line of reasoning presented in Savage and Wainer (2008), arguing for Bayesian methods when detecting suspected terrorists. The context and numbers used here are different, but the approach is identical. Suppose there is a test-taking population of 70,000 examinees—a number not uncommon for a statewide testing program. Furthermore, suppose that 5% of the test-taking population is comprised of cheaters. Finally, suppose that a recently developed statistic, x , shows remarkable promise; it has been shown that using a critical value of X is “99% accurate” both in terms of its sensitivity (i.e., statistical power to detect true cheaters) and its specificity (i.e., true negative rate, or ability to accurately classify noncheaters as noncheaters). This sounds too good to be true, and it probably is. In reality, sensitivity and specificity may not be equal, as one is usually sacrificed in favor of the other; indeed, they are inversely related and we make this assumption for the sake of argument, to demonstrate a best-case scenario.

If the observed value of the statistic, x , for examinee i (x_i) is greater than X (the critical value that promises 99% accuracy), that examinee is flagged as a potential cheater. When x_i is large, the magnitude of this departure from expectation is summarized

through traditional null hypothesis testing: given the individual is *not* a cheater ($\sim C$), what is the probability of observing a value this large (or larger) for x_i ? To answer this, we determine $P(x_i \geq X | \sim C)$ by calculating the area in the right-hand-tail of the null sampling distribution of X . If that probability is sufficiently small (i.e., it falls below an a priori acceptable Type I error rate), then we make the usual inference that the person is probably a cheater; more formally, we say that we reject the null hypothesis. However, that conclusion is not terribly clear to those untrained in statistical inference. More important, it is not a direct inference about the actual subject of interest: What is the probability that the examinee is a cheater? Our evidence thus far indicates that for a noncheater, x_i is an unusual result. But does that imply that examinee i must be a cheater? The answer is: “it might, but we need to know more about the distribution of cheaters to be confident.” The fact is, relying on this traditional frequentist interpretation of p -values alone to flag potential cheaters will result in a lot of incorrect/false positive decisions.

If $N = 70,000$, and 5% are cheaters, then the number of cheaters, N_C , is 3,500 and the number of noncheaters, $N_{\sim C}$, is 66,500. Of the 3,500 cheaters, 99% are accurately flagged by X , resulting in 3,465 correctly identified cheaters. Of the 66,500 noncheaters, only 1% are inaccurately flagged by X , resulting in 665 incorrectly identified cheaters. The total number of flagged individuals is 4,130, or 5.9% of the 70,000 examinees. From this, we can determine the probability that an examinee is a cheater, given she or he was flagged, is $3,465/4,130 = 0.84$, or 84% accuracy. That rate is not terribly low, but it’s not the “99% accuracy” promised. Worse, the complement to this expression is the probability that an examinee is a noncheater, given that she or he was flagged, which is $665/4,130 = 0.16$, a false positive rate of 16%, a far cry from the expected 1% rate.

It gets even worse if the incidence of cheating is far less than the previously assumed 5% marginal proportion. Suppose that the detection statistic, x , and associated critical value, X , can’t detect all cheaters, only those who copy answers (in point of fact, different statistical methods are required for detecting different kinds of cheating, because different kinds of cheating don’t leave the same evidentiary trail). If the proportion of answer copiers in the population is only 1%, then of the 70,000 examinees, N_C is down to 700, whereas the number of noncheaters, $N_{\sim C}$, is up to 69,300. Of the 700 cheaters, 99% are accurately flagged by X , resulting in 693 correctly identified cheaters. Of the 69,300 noncheaters, only 1% are inaccurately flagged by X , resulting in 693 false positives. Thus, the total number of flagged individuals is 1,386, just under 2% of the total, half of which are true cheaters, half of which are not. One would hardly be impressed with a detection statistic with only 50% accuracy, given a flag has occurred, especially if 99% accuracy is expected.

How can this be? What happened to the “99% accuracy?” The problem is that the null hypothesis p -value is providing an accurate representation of the wrong probability. The frequentist p -value is notoriously misunderstood. In this hypothetical example, it is the probability of someone being flagged, given that the person is a noncheater: $P(x_i \geq X | \sim C)$. This probability cannot account for the fact that the marginal proportion of cheaters and noncheaters may be dramatically different; it is furthermore unaffected by *how different* the distribution of X may be for cheaters and noncheaters. As a result, the p -value is a fairly obtuse way of making an inference about something very acute. It is difficult for many to understand “the probability of observing a value this large or larger if the null condition were true.” People will also often erroneously take $1-p$ to be the probability that the null hypothesis is false (e.g., Gelman, 2013), which of course it isn’t: $1-p$ is the probability of a true negative; the probability that a noncheater would

demonstrate a value less than X . That is also useful information, but it likewise does not really tell us anything about whether examinee i is a cheater or not.

Thus, the claim may be true that only “1% of the noncheaters will be erroneously flagged as cheaters,” but this does *not* mean that only 1% of the flagged are noncheaters. This is a typical, and potentially very serious, misinterpretation. This misinterpretation is responsible for a number of erroneous conclusions. For example, as Wainer (2011) has shown, even though mammography is as high as 90% accurate (in terms of sensitivity and specificity), because the incidence rate of breast cancer is so small *compared to* the rate at which women are screened, the false positives far outweigh the true positives. He estimated that the probability of a woman having breast cancer, given a positive mammogram, is a shockingly low 5%, meaning that 95% of women who receive a positive mammogram do not have breast cancer. As we have seen here, the false positive rate *for those flagged* is sure to be much larger than the marginal false positive rate, so claims of “99% accuracy” are potentially very misleading. To estimate the correct probability—that is, the probability that someone is a cheater, given that he or she has been flagged, $P(C|x_i \geq X)$ —one needs to employ Bayesian reasoning.

BAYES’ RULE

Bayes’ Rule, shown in Equation 1, is a formula that deals with conditional probability statements:

$$P(\theta|x) = \frac{P(\theta,x)}{P(x)} = \frac{P(x|\theta)P(\theta)}{P(x)}. \quad (1)$$

It states that the probability of θ , given x (termed the *posterior distribution of θ given x*) is equal to the joint probability of θ and x divided by the marginal probability of x . This expression is equivalent to the probability of x given θ , (often referred to as the *likelihood function of x* , especially in the context of statistical analysis) multiplied by the marginal probability of θ (referred to as the *prior distribution*, as it represents the distribution of θ without regard to x) divided by the marginal probability of x . This result seems benign in and of itself. It is, in fact, an effective way of solving a number of important conditional probability problems, such as, “what are the chances someone is a terrorist, given some suspicious behavior?” (Savage & Wainer, 2008), or, “what is the probability a women has breast cancer, given a positive mammogram result?” (Wainer, 2011).

However, the implementation of Bayes’ Rule for statistical data analysis in the social sciences can be contentious. Using the notation above, the values of θ may be the parameters of a statistical model, or some other estimand of interest, and the values in x are the observed data. In terms of cheating detection, θ could be characterized as “examinee i is a cheater” (which will henceforth be denoted with a “C”), and x could be characterized as “observing a value for our cheating detection statistic, x_i , greater than or equal to X .” Equation 2 reframes Bayes’ Rule in that context:

$$P(C|x_i \geq X) = \frac{P(x_i \geq X|C)P(C)}{P(x_i \geq X)}. \quad (2)$$

A few definitions follow, leading to an expression for how to solve Equation 2 in practical terms. First, the denominator of this formula, $P(x_i \geq X)$, represents

the marginal distribution of extreme X values; that is, it is the proportion of all observed x_i values which are greater than or equal to X . If one has established a detection threshold via an acceptable marginal Type I error rate (i.e., the α level) and associated critical value, then $P(x_i \geq X)$ is simply the proportion of observed values that meets or exceeds that threshold; this is easily determined from the data distribution. In the motivating example, this probability came from counting up the total number of flagged examinees. (The example used ratios of frequencies, both of which would have been divided by a constant, N , to make them probabilities, so the N was left out.)

The numerator of the formula contains a product of two terms that are not directly observed, but may be estimated based on weak assumptions. $P(x_i \geq X|C)$ is the likelihood of observing a value of x_i greater than or equal to X , given examinee i is a cheater. In this context the likelihood represents the power of the statistical test, the probability that a flag is obtained for a true cheater. $P(C)$ is the *prior* probability for cheaters, which represents the proportion of cheaters in the population. This may be known a priori, based on previous experience, or the analyst may make a reasoned estimate of its value; regardless, the impact of the prior tends to be the most contentious issue in Bayesian analysis, so this topic will be addressed directly in a subsequent section of this chapter. For now, we continue on with estimating the numerator of the posterior density expression. To arrive at the estimate, we first note that the marginal density, $P(x_i \geq X)$, by definition is equal to:

$$\begin{aligned} P(x_i \geq X) &= \sum_{\theta} P(x_i \geq X | \theta)P(\theta) \\ &= P(x_i \geq X | C)P(C) + P(x_i \geq X | \sim C)P(\sim C). \end{aligned} \tag{3}$$

In Equation 3, θ represents any unknown parameter and all of its possible values; in this case, those values are “examinee i is a cheater,” represented by C , and “examinee i is a noncheater,” represented by $\sim C$. Those are the only two possible values for the parameter of interest. Solving for the posterior’s numerator, $P(x_i \geq X|C)P(C)$, we obtain the Equation 4:

$$P(x_i \geq X | C)P(C) = P(x_i \geq X) - P(x_i \geq X | \sim C)P(\sim C). \tag{4}$$

In the motivating example, the numerator was estimated by counting up the number of correctly flagged cheating examinees (as previously stated, that example used ratios of frequencies, not probabilities, but operated on the same principles).

There is good news and bad news here for the analyst who is unconvinced that the Bayesian approach is preferred. The good news is that there are some familiar terms that can be estimated directly from the data. As mentioned, $P(x_i \geq X)$ is simply the marginal proportion of x_i values greater than or equal to X . $P(x_i \geq X | \sim C)$ is the probability that $x_i \geq X$, given examinee i is a noncheater. This expression is the well-known p -value from a null hypothesis test. The potential bad news is that $P(\sim C)$ and $P(C)$, the prior probabilities of noncheating and cheating, respectively, cannot be directly estimated from the data. In the motivating example, we posited the marginal incidence of cheating was 5% or 1%. In practice, this probability would have to be estimated. Informed judgment or previous experience and data may supply reasonable estimates, but the analyst only has to estimate one of these: because $P(\sim C) + P(C) = 1$, every examinee is by definition either a noncheater or a cheater. Assuming for now an estimate can be

obtained, we return to the solution for the posterior probability of cheating, given the data, substituting into the formula the result for its numerator:

$$P(C|x_i \geq X) = \frac{P(x_i \geq X|C)P(C)}{P(x_i \geq X)} \quad (5)$$

$$= \frac{P(x_i \geq X) - P(x_i \geq X|\sim C)P(\sim C)}{P(x_i \geq X)} \quad (6)$$

$$= \frac{P(x_i \geq X)}{P(x_i \geq X)} - \frac{P(x_i \geq X|\sim C)P(\sim C)}{P(x_i \geq X)} \quad (7)$$

$$= 1 - \frac{P(x_i \geq X|\sim C)P(\sim C)}{P(x_i \geq X)} \quad (8)$$

$$= 1 - P(\sim C|x_i \geq X). \quad (9)$$

Equation 9 is just another application of Bayes' Rule. This result is mathematically consistent and logical because $P(C|x_i \geq X) + P(\sim C|x_i \geq X) = 1$. For any given value of x_i greater than or equal to X , examinee i must either be a cheater or a noncheater.

Thus, there is a relatively simple formula to solve now, and all it requires is to estimate (1) the marginal proportion of x_i values greater than or equal to X , (2) the null hypothesis p -value associated with this threshold, and (3) a reasonable estimate of the proportion of cheaters (and, correspondingly, the proportion of noncheaters) in the population. We can then turn a frequentist p -value into a Posterior Probability of Cheating (PPoC) as follows:

$$PPoC = 1 - \frac{P(x_i \geq X|\sim C)P(\sim C)}{P(x_i \geq X)} \quad (10)$$

$$= 1 - \frac{p\text{-value} \times P(\text{non-cheater})}{P(\text{data above threshold})}, \quad (11)$$

where the p -value in the numerator of Equation 11 is the null hypothesis p -value. That value is multiplied by an estimate of the proportion of the population who are not cheaters, and the result is divided by the proportion of examinees who demonstrate x_i values above the threshold, X . This is the posterior probability of being a *noncheater*, given the data; subtracting that value from one provides the PPoC. A useful feature is that from these two probabilities – $P(C|x_i \geq X)$ and $P(\sim C|x_i \geq X)$ – one can construct Bayes factors (Jeffreys, 1960), which are odds ratios, in this case, $P(C|x_i \geq X)/P(\sim C|x_i \geq X)$. This simple transformation conveys the probability that a flag indicates a true cheater on the odds scale.

METHOD

The benefits of calculating the PPoC, as opposed to relying on traditional p -values, is demonstrated by means of a series of analytic examples. This analysis does not focus on particular cheating statistics, but rather generalizes detection to any such statistic. These analytic examples demonstrate the problem with null hypothesis testing and “statistical significance” for very low-incidence events.

First, consider a hypothetical cheating detection statistic, x . Under the null hypothesis (H_0), the sampling distribution of x is standard normal: $x|C \sim N(0, 1)$. Further, assume that 1% of the population is comprised of cheaters; thus $P(C) = 0.01$ and $P(\sim C) = 0.99$. Sixteen conditions were created to represent various cheating detection scenarios: four detection threshold values, X_c , representing increasing specificity (threshold X_c values = 2, 3, 4, 5), crossed with four expected values for the sampling distribution of x for cheaters. Thus, $x|C \sim N(\mu, 1)$, with $\mu = 2, 3, 4, 5$. For each condition, a unique mixture distribution of normal distributions is constructed and the PPoC is calculated by using true population values for three proportions: $P(x_i \geq X_c)$, the marginal proportion of x_i values greater than or equal to X ; $P(x_i \geq X_c|\sim C)$, the null hypothesis p -value associated with this threshold; and $P(\sim C)$, the true proportion of noncheaters, equal to 0.99. Because all supplied values are analytically derived, the resulting PPoC values are true. As previously stated, with real data $P(x_i \geq X_c)$ and $P(x_i \geq X_c|\sim C)$ could be calculated directly, but $P(\sim C)$ would, in practice, have to be estimated. As such, the influence of correctly or incorrectly specifying $P(\sim C)$ is then further considered.

RESULTS AND DISCUSSION

Tables 18.1–18.3 contain true values for the likelihood values, $P(x_i \geq X|C)$, marginal distributions for $x_i \geq X$, $P(x_i \geq X)$, and the PPoC values, $P(C|x_i \geq X)$, respectively. Each table contains these values for the 4×4 crossed conditions. For each of the four X thresholds, the corresponding column is labeled by its null hypothesis alpha level (i.e., the area in the right-hand-tail of the null distribution, or the minimally sufficient p -value to reject H_0). The true likelihood values and marginal distributions of $x_i \geq X$ are included in Tables 18.1 and 18.2 as reference points. PPoC values in Table 18.3 are of primary interest for inference making. These can be constructed by multiplying a corresponding likelihood by 0.01 (the prior probability of cheating) and dividing by $P(x_i \geq X)$.

The likelihood values in Table 18.1 represent the statistical power of each null hypothesis test. That is, they represent the probability that the threshold will correctly identify a true cheater. The likelihood values show that when the X threshold is equal to the expected value of the cheaters' distribution, there is only a 50% chance that $x_i \geq X$ for cheaters (i.e., the threshold is the median of the score distribution for cheaters). When the expected value exceeds the threshold, the power is considerably higher. Conversely, when the threshold is below the expected value, the probability that $x_i \geq X$ for cheaters is quite low. These changes are noteworthy because their magnitudes are

Table 18.1 True Likelihood (Power) Values, $P(x_i \geq X|C)$, by Condition

μ for cheaters	Detection threshold X			
	$X = 2, p\text{-value}$ $\leq 2.275 \times 10^{-2}$	$X = 3, p\text{-value}$ $\leq 1.350 \times 10^{-3}$	$X = 4, p\text{-value}$ $\leq 3.167 \times 10^{-5}$	$X = 5, p\text{-value}$ $\leq 2.867 \times 10^{-7}$
2	0.500	0.159	0.023	0.001
3	0.841	0.500	0.159	0.023
4	0.977	0.841	0.500	0.159
5	0.999	0.977	0.841	0.500

Table 18.2 True Marginal Distributions of $x_i \geq X_c$, $P(x_i \geq X)$, by Condition

μ for cheaters	Detection threshold for x			
	$X = 2, p\text{-value}$ $\leq 2.275 \times 10^{-2}$	$X = 3, p\text{-value}$ $\leq 1.350 \times 10^{-3}$	$X = 4, p\text{-value}$ $\leq 3.167 \times 10^{-5}$	$X = 5, p\text{-value}$ $\leq 2.867 \times 10^{-7}$
2	0.028	0.003	<0.001	<0.001
3	0.031	0.006	0.002	<0.001
4	0.032	0.010	0.005	0.002
5	0.033	0.011	0.008	0.005

Table 18.3 True PPoC Values, $P(C|x_i \geq X_c)$, by Condition

μ for cheaters	Detection threshold for x			
	$X = 2, p\text{-value}$ $\leq 2.275 \times 10^{-2}$	$X = 3, p\text{-value}$ $\leq 1.350 \times 10^{-3}$	$X = 4, p\text{-value}$ $\leq 3.167 \times 10^{-5}$	$X = 5, p\text{-value}$ $\leq 2.867 \times 10^{-7}$
2	0.18	0.54	0.88	0.98
3	0.27	0.79	0.98	0.99
4	0.30	0.86	0.99	0.99
5	0.31	0.88	0.99	0.99

considerably larger than the corresponding changes in p -values. Type I error and power are always directly related, though not linearly.

The $P(x_i \geq X)$ values in Table 18.2 are consistent with expectation. As μ for cheaters increases, so does the marginal proportion of extreme x values. Conversely, as the detection threshold is increased, the marginal proportion of extreme x values decreases. The fact that these proportions are so small (ranging from practically zero to no higher than 3.25%) is important for explaining the PPoC values in Table 18.3.

The PPoC values are presented in Table 18.3 and illustrated in Figures 18.1 and 18.2. Figure 18.1 illustrates how the PPoC is calculated, using the example of detection threshold $X_c = 2$, and $\mu|C = 5$. In this example, even though the likelihood, $P(x_i \geq X_c | C)$, is 0.999, only 1% of the total population is comprised of cheaters, $P(C) = 0.01$. With the threshold set at $X_c = 2$, there are 2.275 times more incorrectly flagged noncheaters than there are correctly flagged cheaters. Thus, the PPoC is only 0.31. Figure 18.2 shows this same relationship, with the area around the flagging region magnified to enhance the details of the two distributions.

As evidenced by these results, there can be a considerable difference between finding a statistically significant p -value (i.e., because $x_i \geq X_c$) from a null hypothesis test, and having a reasonably high probability of correctly identifying a cheater. For low-incidence events such as these, if the detection threshold is relatively low (e.g., only two standard errors above the mean), the probability that examinee i is a cheater, given $x_i \geq X$, is considerably lower than the probability that he or she is a noncheater. For any value of μ , as the detection threshold is increased, so does the PPoC. The increase

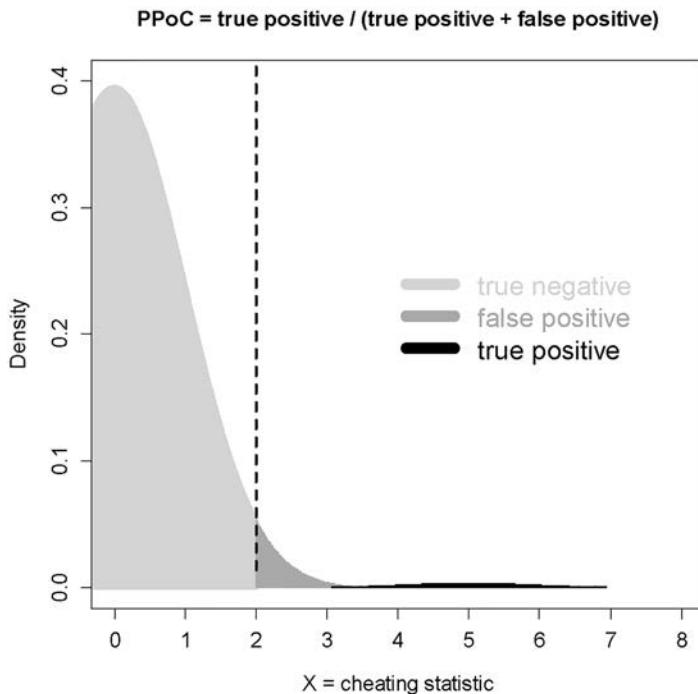


Figure 18.1 The PPoC, $P(C|x_i \geq X)$, is the Marginal Proportion of true positives, $P(x_i \geq X|C) P(C)$, divided by the probability of being flagged, $P(x_i \geq X)$. This is equivalent to the number of true positives divided by the number of flagged examinees. In this example, $X=2$ and $m|C=5$.

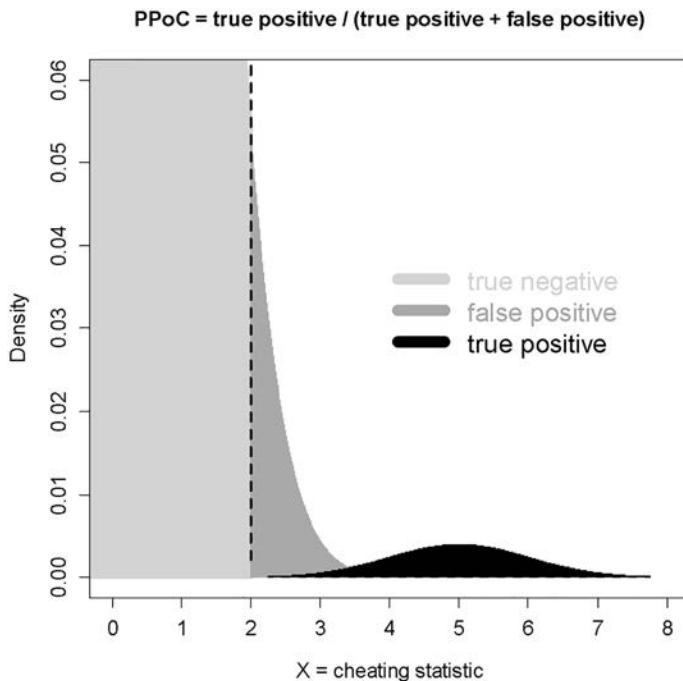


Figure 18.2 Magnification of the Flagging Region Shown in Figure 18.1 to Enhance Detail

in PPoC is larger for increases in the threshold than it is for increases in μ . When the detection threshold is set to $X = 5$ (five standard errors above the mean), the probability that examinee i is a cheater, given $x_i \geq X$, is nearly one. Of course, such a large threshold may very well sacrifice sensitivity for the sake of specificity; in this example, thresholds of three and four are higher-powered but still have fairly large PPoC values.

This analytic exercise is instructive, but it assumes one knows the prior probability of cheating, which is unlikely to be true in practice. How does one proceed without knowing that value? One answer is to consider a range of plausible priors, and see how that choice influences the posterior inferences. If the inferences are greatly influenced by the priors, then the analyst knows that the information in the likelihood is not very strong, whereas small changes in the posterior probabilities mean that the prior's influence is not very strong. To demonstrate these differences, we extend the analytic demonstration to a case that approximates reality, whereby $P(C)$ is not known and must be estimated.

Table 18.4 contains a series of PPoC values derived by iteratively changing the value of $P(C)$. In this example, the cheater distribution expected value is held constant ($\mu = 5$, corresponding to the last row in Table 18.3), but the choices for $P(C)$, ranging from 0.001 to 0.5, are completely crossed with the four threshold conditions. We see that when the correct prior is used (in the highlighted row), we obviously obtain the correct PPoC value. But another interesting pattern is clear: The estimate of PPoC is much less influenced by the choice of prior when the true PPoC value is closer to one. For the $X_c = 2$ threshold, PPoC estimates ranged from 0.30 for $P(C) = 0.001$ up to 0.65 for $P(C) = 0.5$. Estimating the percentage of cheaters at 50% is almost certain to be an overestimate (one would hope), but these extreme cases give us insight into the stability of the PPoC. Conversely, for the $X = 3$ threshold, PPoC estimates ranged from 0.88 for $P(C) = 0.001$ up to 0.94 for $P(C) = 0.5$, a range of only 0.06. For the $X_c = 3$ and $X_c = 4$ thresholds, the range of PPoC estimates is practically zero.

Table 18.3 and Figure 18.1 helped establish that PPoC values are closer to one when the detection threshold moves away from the null distribution (and, to a lesser extent, when the distribution of x for cheaters moves further away from the null distribution). The reason why the threshold is more influential than the expected value of the cheater

Table 18.4 Estimated Bayesian PPoC Values by Detection Threshold (X) (and Associated Null P -Value) Crossed With Various Prior Specifications for Cheating Prevalence

Prior specification for Marginal Probability of Cheating, $P(C)$	$X_c = 2, p\text{-value}$ $\leq 2.275 \times 10^{-2}$	$X_c = 3, p\text{-value}$ $\leq 1.350 \times 10^{-3}$	$X_c = 4, p\text{-value}$ $\leq 3.167 \times 10^{-5}$	$X_c = 5, p\text{-value}$ $\leq 2.867 \times 10^{-7}$
0.001	0.301	0.879	0.996	0.999
0.01	0.307	0.880	0.996	0.999
0.05	0.335	0.885	0.996	0.999
0.1	0.370	0.891	0.997	0.999
0.2	0.440	0.903	0.997	0.999
0.3	0.510	0.915	0.997	0.999
0.4	0.580	0.927	0.998	0.999
0.5	0.650	0.939	0.998	0.999

Note: Calculations are computed here for the true cheating distribution with $\mu = 5$. The highlighted row, $P(C) = 0.01$, is correctly specified.

distribution is due to the very large discrepancy in the prior probabilities of cheating and not cheating. When 99% of the frequency distribution is noncheaters, the only way to confidently say a flag probably indicates a cheater is to make the threshold so high that practically all noncheaters would be excluded. At that point, within the distribution of examinees for whom $x_i \geq X$, there are more cheaters by far than noncheaters. It is precisely this kind of inference making that the Bayesian paradigm encourages and that the frequentist reliance on *p*-values completely ignores.

CONCLUSION

In 1976, George Box famously stated “all models are wrong” (p. 792) but may nonetheless be useful, especially when parsimonious. Bayesian methods persist because they are eminently useful, and can be applied parsimoniously, even if they are sometimes wrong. An important feature of these procedures is they come with a built-in way of evaluating how wrong they are. If one is concerned that the prior may be too influential, one can consider other, competing priors and see how much, if at all, the answer changes. There are three great benefits to this way of thinking: (1) one attends to the continuous nature of probability, as opposed to focusing on yes/no decisions from the null hypothesis (Gelman, 2011, 2013); (2) one has a practical way to construct meaningful credible intervals for the PPoC; and (3) ultimately, one estimates the probability that is of actual interest, the probability that someone is a cheater, given the observed value on a statistic of interest.

As the final example in Table 18.4 demonstrates, the choice of a prior may be fairly influential on the estimate of the Bayesian posterior probability. This is often cited as a concern regarding the Bayesian paradigm, but it in fact allows the Bayesian analyst to treat the prior as just another source of information, not unlike the data themselves. Initially, there may be little evidence to suggest an obvious value for $P(C)$. An informed guess could be the starting point, for example, cheating is probably less prevalent than not cheating, so start with 0.5 or lower for $P(C)$. Furthermore, the Bayesian paradigm actually invites the analyst to consider a range of possible values for the prior probability, as was demonstrated in Table 18.4. Evaluating the influence of the prior provides insight into the nature of the posterior, as with the comparison of ranges in posterior probabilities from the $X_c = 2$ column in Table 18.4 with the same range from the $X_c = 5$ column. With a relatively low threshold for detection, there is greater uncertainty in the PPoC as a function of the prior information, but as the threshold increases, the prior probability has less and less influence, because the likelihood dominates the equation. This enlightening information is provided by considering multiple priors, while the traditional *p*-value remains a constant for a given threshold.

One can not only evaluate a range of priors and examine their influence, one can also change these values as new information comes to light. For example, perhaps a series of investigations turns up evidence that certain flagged examinees were cheaters and others were not. If subsequent data analysis and experience demonstrate that the prior probability is surely somewhere between 0.01 and 0.10, for example, then posterior inferences can be that much more influenced by this information. Priors can be updated accordingly and posteriors may be recalculated to make stronger inferences. Traditional null hypothesis testing does not consider the distribution of the actual parameter of interest (in this case, the prevalence of cheating), so it cannot benefit from such updated information.

Although the context of this discussion is cheating detection, the arguments contained herein really apply to all statistical diagnostic decision-making practices, particularly those dealing with low-incidence phenomena. Moreover, the reasons for adopting a Bayesian approach in a wide variety of parameter estimation contexts are merely extensions of this line of reasoning. In recent years, Bayesian statistics have enjoyed a renaissance, as pragmatists have seized upon techniques like Markov Chain Monte Carlo as a means to estimate parameters when traditional maximum likelihood might not work. However, many of these analysts nonetheless limit the role of the prior to the extent possible. In the words of L. J. Savage (1961), they attempt to “make the Bayesian omelet without breaking the Bayesian eggs” (p. 575). We go a step further to say that Bayesian methods are both theoretically sound and of practical advantage because they incorporate prior information. We say: “Break the eggs, make the omelet, and use all the information available to make as informed an inference as possible.” In closing, we began with a quote from Bertrand Russell, and will conclude with another: “In all affairs it’s a healthy thing now and then to hang a question mark on the things you have long taken for granted” (unknown source). We suggest that it is time to hang a question mark on null hypothesis testing and our dogged reliance on estimating the wrong probability.

REFERENCES

- Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, 69, 44–49.
- Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71, 791–799.
- Cannell, J. J. (1989). *How public educators cheat on standardized achievement tests: The “Lake Wobegon” report*. Albuquerque, NM: Friends for Education.
- Cizek, G. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gelman, A. (2011). Induction and deduction in Bayesian data analysis. *Rationality, Markets, and Morals*, 2, 67–68.
- Gelman, A. (2013). *P* values and statistical practice. *Epidemiology*, 24(1), 69–72.
- Jeffreys, H. (1960). *Theory of probability* (3rd Ed.). Oxford: Clarendon.
- Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, 8, 261–272.
- Savage, J. L. (1961). The foundations of statistical inference reconsidered. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, 575–586.
- Savage, S., & Wainer, H. (2008). Until proven guilty: False positives and the war on terror. *Chance*, 21(1), 59–62.
- Thiessen, B. (2007). Case study—Policies to address educator cheating. Retrieved from: <http://homepage.mac.com/bradthiessen/pubs/format.pdf>
- Wainer, H. (2011). How should we screen for breast cancer? *Significance*, 8(1), 28–30.
- Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement*, 21, 307–320.
- Wollack, J. A. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement*, 40, 189–205.

19

WHEN NUMBERS ARE NOT ENOUGH Collection and Use of Collateral Evidence to Assess the Ethics and Professionalism of Examinees Suspected of Test Fraud

Marc J. Weinstein

INTRODUCTION

This chapter provides an overview of methods used to identify, collect and preserve information and evidence, collateral to examination results, for the purpose of identifying examinees whose conduct falls short of a test sponsor's standards for ethics, character and professionalism.¹ As an overview, the chapter does not address in detail all of the potential sources of collateral evidence a test sponsor may collect, or provide examples of the utility of every type of collateral evidence identified herein. The chapter is merely intended as a starting point to consider these issues and, through the use of the hypothetical investigation scenarios described herein, to promote discussion among professionals in the testing community about best practices for identifying, collecting and preserving collateral evidence of test fraud.

THE NEED TO COLLECT EVIDENCE COLLATERAL TO EXAM RESULTS

There appears to be consensus within the test sponsor community that the detection and investigation of potential test fraud should be primarily focused on the validity of test results, not the character of examinees. In other words, test sponsors should only be concerned with determining whether there is something about an examinee's test result that makes it an invalid measure of the examinee's knowledge on the subject matter of the test. In this way, the test sponsor is not concerned with making any judgment about the ethics, morality or character of the examinee, or proving that the examinee engaged in any specific wrongful conduct that enabled him or her to achieve an invalid score. Indeed, reliable evidence of test fraud derived entirely from exam results (i.e., unusual gains, extreme similarity of responses, wrong to right answer changes and

other types of score aberrance) can provide a test sponsor with a sufficient basis to cancel an examinee's score.

However, a purely statistical approach does not serve the needs of all test sponsors. Some test sponsors prefer to obtain and analyze information outside of the test results (hereafter, *collateral evidence*) prior to taking any action with respect to an examinee suspected of cheating or test fraud. This is especially true where ethics, good character and professionalism are essential components of the core values of the testing organization and necessary qualifications for the examinee to obtain or maintain the credential sought.

Test sponsors that place a high value on the ethics, good character and professionalism of their examinees include, but are not limited to, medical specialty certifying boards, financial industry regulators and certification organizations, professional licensing agencies and higher education organizations that administer examinations for admission to professional degree programs. Over the past several years, many such test sponsors have become more outspoken in reminding examinees that the ethics, character and professionalism components of their credentials apply to their examinations. For example, the American Board of Medical Specialties, which represents 24 medical specialty boards, posted the following statement on its website in 2012:

Patients and their families place enormous trust in physicians and other medical professionals. This trust must be continuously earned and cultivated. As the gold standard among medical credentials, Certification by one of the 24 Member Boards of the American Board of Medical Specialties (ABMS) is an indicator of quality and professionalism that patients and their families have come to rely on.

Upholding public trust is one reason ABMS takes exam security very seriously. Secure exams are one important step in the lifelong ABMS-led learning process that ensures certified physicians meet high standards for the knowledge and skills they bring to their patients.

Exam security and meaningful certification require multi-faceted approaches that include continuous improvement and vigilance. ABMS supports its 24 Member Boards as they assure the highest quality exam security through:

- identifying cheating with the latest monitoring; technologies;
- routinely changing test questions;
- strengthening the security of testing sites;
- auditing exam security; and
- designing a rigorous certification process.

It should be made abundantly clear that recalling and sharing questions from exams violates exam security, professional ethics, and patient trust in the medical profession. When it happens, the practice should be addressed swiftly and decisively. Whether someone is providing or using test questions, ABMS Member Boards enforce sanctions that may include permanent barring from certification, and/or prosecution for copyright violation.

Physicians who rely on recalled questions to prepare for certification exams should know that doing so not only violates the policies of the ABMS Member Boards but utilizing recalled questions is an unreliable way to prepare for such exams. Member Boards routinely change test questions, and a high percentage of shared questions and answers are recalled incorrectly. In contrast, ABMS Member Boards are offering more ways to help residents legitimately and ethically prepare for exams.

The public should also feel assured that certification requires more than passing a single exam. In fact, every ABMS Member Board now requires physicians to demonstrate their knowledge, skills, professionalism and ethics throughout their careers to maintain their certification.

(American Board of Medical Specialties, 2012, p. 1, para. 6)

In light of the foregoing statement by the ABMS, to merely cancel an examinee's score in response to a statistical analysis finding probable test fraud would not be sufficient to uphold the organization's values. Rather, an ABMS member board must determine whether there is something about the examinee's conduct in relation to the exam that suggests that he or she lacks the ethics, good character or professionalism to become a board certified physician or to maintain board certification.

Thus, in situations where a test sponsor detects potential or likely test fraud based upon data derived entirely from exam results, it is critical to such test sponsors to determine not only whether an examinee's score is not a reliable measure of the person's true knowledge of the subject matter tested but also whether the person engaged in conduct that fails to meet the high standards of ethics, character and professionalism required by the test sponsor. In cases such as this, the test sponsor must identify, collect and preserve reliable collateral evidence to determine whether an examinee has personally engaged in conduct that falls short of the ethics, character or professionalism required by the sponsoring organization. To simply invalidate an examinee's score and allow an examinee to retest after the test sponsor has detected score aberrance indicating likely test fraud would do little to uphold a core value of the testing organization.

COLLECTION AND RETENTION OF COLLATERAL EVIDENCE AS PART OF THE TEST SPONSOR'S VERTICALLY INTEGRATED TEST SECURITY PROGRAM

Sponsors of secure examinations should have a comprehensive test security program that is vertically integrated into all departments of the sponsor's organization. Although organizational structure varies from sponsor to sponsor, many test sponsors have the following departments: human resources, information technology, test development, examinee registration and records, test administration, psychometrics, communications, finance, legal and executive. Vertical integration of a test security program requires the test sponsor to consider and include all components of the organization that impact exam security and ensure that the persons employed in those departments understand their respective roles in preventing, detecting and responding to test security incidents, and how they must interface with responsible persons in other departments for the purpose of ensuring test security. A vertically integrated comprehensive test security program will identify the exam security roles and responsibilities of all persons within each department of the organization and provide a workflow or series of interrelated workflows for preventing, detecting and responding to test security incidents.²

The author assumes for purposes of this chapter that the test sponsor has already designed and implemented a comprehensive, effective and vertically integrated test security program. A central element of an effective test security program is a comprehensive examinee agreement that clearly states the rights and responsibilities of the test sponsor and the examinee. Indeed, an examinee agreement is one of the most important building blocks of an effective test fraud investigation. Boiled down to its essential

characteristics, an effective examinee agreement must clearly establish five elements: (1) permissible and impermissible examinee conduct prior to, during and after the exam; (2) the right of the test sponsor to use statistical data analyses and other methods to detect cheating; (3) the right of the test sponsor to further investigate score aberrance and any suspected breaches of test security; (4) the obligation of examinees to cooperate in any investigation by the test sponsor; and (5) the potential consequences of a finding of score invalidity and/or any violation of the examinee agreement and the organizational process for determining and imposing such consequences.

Vigorously adhering to a comprehensive test security program and enforcing the terms of the test sponsor's examinee agreement will ensure that the test sponsor is employing all reasonable measures available to prevent and detect test fraud, including but not limited to statistical data analyses. Indeed, it is presumed that the results of such data analyses by a test sponsor could serve as the starting point for the investigation strategies set forth in this chapter.

THREE DISTINCT TIME PERIODS TO COLLECT EVIDENCE

For every examination, there are three distinct time periods when a test sponsor can identify, collect and preserve collateral evidence that may become relevant to an investigation of the examinee's conduct with respect to an exam: before, during and after the exam. Each of these time periods offers an opportunity to collect different types of collateral evidence that a test sponsor may find critical when conducting an investigation.

Although it may be helpful to identify and collect collateral evidence in response to the detection of possible test fraud, it is not nearly as effective as implementing procedures that automatically provide for the collection of such evidence as part of the test sponsor's exam registration and administration processes. By integrating and automating the collection of collateral information in this way, the test sponsor will collect a myriad of useful evidence that investigators can immediately utilize following a statistical finding suggesting that the examinee's score is not valid.

Finally, it must be acknowledged that not all test sponsors have exam registration systems or administration practices that enable the collection of all of the suggested collateral evidence. From an evidence collection perspective, there are obviously significant advantages to computer-based testing; however, much of the collateral evidence identified in the following sections can be collected and retained even for paper-based exam administrations.³

Before the Examination

Before the exam, test sponsors must collect biographical, personal and transactional information about the candidate that will enable them to analyze the examinee's score in a meaningful context if the examinee is suspected of test fraud or aberrance is later detected in the examinee's test score. Ideally, the basic information that should be collected and retained by test sponsors concerned about the ethics, character and professionalism of their examinees would include, but not be limited to, the following:

- name, alias and all prior names;
- date and place of birth;
- Social Security number;

- current color photograph of the examinee;
- color copy of driver's license and/or passport;
- present physical address, address history for a period of at least 3 years;
- all email addresses presently and previously used by the examinee;
- educational history, including but not limited to academic achievement at each institution;
- prior standardized examinations taken and scores achieved;
- professional licenses held;
- employment history for a period of at least 3 years;
- prior criminal and disciplinary history, including any history of academic dishonesty;
- identification of family members who have taken or registered to take the same examination and the dates of those examinations;
- exam score history for the exam at issue;
- complete registration and cancellation history for the exam at issue;
- exam score history for all practice examinations taken by the examinee;
- payment information for all methods of payments to the test sponsor;
- test center selection(s) and proximity to current residence;
- all IP addresses used to access the test sponsor's website; and
- all social media sites used by the examinee and related user names.

Each of these pieces of examinee information, although collateral to exam results, may be useful in its own right to the test sponsor for a variety of purposes. However, the information will become even more valuable if the test sponsor is searching for collateral evidence of test fraud following a determination of aberrance in the examinee's score or some other reason to suspect cheating or related conduct in violation of the examinee agreement. The test sponsor should also require examinees to submit changes and additions to most of the above information on an ongoing basis following its initial submission, to the extent such information is known by the examinee. For example, although it makes sense to require examinees to submit changes to their names, addresses, criminal history and other personal information that they would have reason to be aware of, it would be unreasonable to expect examinees to know when their IP addresses change. Finally, the test sponsor's examinee agreement should make clear that the provision of false, misleading or incomplete information at any stage of the registration or examination process is a violation of the agreement and constitutes grounds to impose sanctions against the examinee.

The collection of a current photograph and valid driver's license or passport copy as early as possible in the registration process is a defensive measure designed to discourage the use of a proxy by an examinee. Proxy schemes are much less likely to succeed when the examinee is required to produce a valid driver's license and have his or her photograph taken on the day of the examination—which may be several weeks or even months after having submitted the photograph and driver's license for registration. If, however, the examinee produces his or her driver's license or passport for the first time on the day of the exam, and has never previously tested or submitted a photograph, finger print or palm vein scan to the test sponsor, any person, using a fake identification, could present herself at the test center as the examinee. The test sponsor's collection and retention of examinee photographs and identifications submitted at the time of registration is therefore critical to enable the test sponsor to compare exam day photographs and identifications presented by examinees.

During the Examination

During the exam, there are numerous opportunities to collect collateral evidence of an examinee's conduct that may assist a test sponsor if the examinee is suspected of test fraud or aberrance is later detected in the examinee's test score. To begin with, the test sponsor should use an exam-day admittance process designed not only to verify the examinee's identity but also to collect collateral evidence. To gain admittance to an exam, the examinee should be required to present the same valid driver's license, passport or other official government identification that was submitted at the time of registration. The identification should be scanned and copied and the person's photograph should be taken. In addition, either a palm vein scan or fingerprint should be taken for comparison and readmittance to the exam following breaks. The entire exam admittance and screening process should be audio and video recorded. Finally, the examinee should be required to sign her name when first admitted and for readmittance after each break.

Examinees should be required to empty their pockets, take off all outerwear (i.e., jackets, sweaters, sweat shirts, hats, scarves, etc.) and deposit all of their personal belongings, including all electronic devices, into a locker outside of the screened entry area. The locker area itself should be under constant video and audio surveillance. Each examinee should be asked to pull his pockets inside out to show that they are empty and, where permissible, examinees should be checked with a metal detecting wand. A seating chart should be maintained for all examinees admitted into the testing room. If the exam is being administered at the same time as other exams of a shorter duration, and the people taking the other exams will change through the course of the exam at issue, the seating chart should be updated for each time frame when changes occur. As alluded to above, examinees should be required to sign out for each break and subjected to the identical identification, biometric and physical screening before readmittance following each break. This is another measure designed to defeat the substitution of an examinee with a proxy following a break. If an examinee cannot establish a biometric match upon return from a break, the proctor must take action to establish and confirm the identity of the examinee, or simply terminate the exam, depending upon the sponsor's policies.

In addition, examinees should be video- and audiotaped while taking their exam, and proctors should observe examinees 100% of the time that they are taking their exam. The proctor must immediately investigate any unusual behavior or observed communication by examinees. Unusual behavior can range from talking to another examinee, to pulling a piece of paper out of a pocket, tapping on a desk, or frequently leaving the testing room to go to the bathroom. These and any other unusual behaviors must be immediately investigated, recorded and reported to the test sponsor.

If a proctor or another examinee observes unusual conduct or testing rule violations, the test sponsor has a limited amount of time within which to conduct a truncated investigation that may prove critical in determining whether the examinee engaged in intentional misconduct for the purpose of gaining or assisting another person in gaining an unearned advantage on the exam. There is certainly room for debate concerning the most effective means of accomplishing such a truncated investigation in the middle of an exam. Indeed, some test sponsors may be flatly opposed to interrupting an exam for purposes of such an investigation. However, an investigation undertaken within moments of the potential misconduct can be incredibly effective in distinguishing between intentional misconduct and mere negligence in following the exam rules by the examinee.

Hypothetical Mid-Exam Irregularity Investigation

Consider a scenario where a test sponsor's rules prohibit examinees from accessing or using a mobile phone for the entire duration of an exam, including during breaks (and all test sponsors should have such a rule). For purposes of this hypothetical scenario, a proctor observed an examinee take a mobile phone out of his or her locker and walk out the front door of the test center. In a circumstance such as this, the proctor would stop the clock on the examination for that examinee. The proctor would then contact the test sponsor to report the incident and get further guidance on how to proceed before readmitting the examinee to resume his or her exam. This author would advise the proctor to place the examinee on a telephone call with the test sponsor's exam security investigator who should, after reminding the examinee of the agreement he or she entered and the rules against accessing and/or using a mobile phone during the exam, and ask whether the examinee accessed prohibited materials (including the Internet) or communicated with anyone. If so, the examinee should be asked about who he or she communicated with, why he or she did so, and to describe the subject of the communications. The investigator should further ask for the name and telephone number of the person with whom the examinee communicated with and advise that a proctor is going to inspect the mobile device for evidence of online activity, calls and texts within the time frame of the exam. Then the investigator should have the proctor look at the mobile phone to inspect it for these purposes and verify the information provided by the examinee. The proctor should take digital photographs of the mobile phone displaying the relevant screens (i.e., Web browser pages, call history and text messages) to ensure that the evidence is preserved. If the exam is a paper-and-pencil examination, the answer sheet and any allowed scrap paper should be photocopied in its current state before allowing the examinee to reenter the exam. This entire process should take less than 15 minutes.

Continuing with the hypothetical scenario, suppose that the examinee explains that he or she was concerned because of a sick child in another person's care during the exam and wanted to check on the health of the child and the information provided regarding the person with whom he or she communicated is consistent with what is found in the phone, and there is no evidence of Internet usage or inappropriate text messages. In such a situation, the test sponsor may decide to allow the examinee to resume the exam. If the examinee refuses to cooperate with any of the previously described steps or provides information about the use of the mobile device that is inconsistent with the evidence found during the inspection of the device, the test sponsor may simply advise the proctor to terminate the exam. However, even under the latter circumstances, experience militates in favor of allowing the examinee to complete the exam so that the test sponsor can analyze the examinee's answer patterns following readmittance to the exam, to determine whether the data further corroborates or disproves the evidence gathered in the mid-exam investigation. Thus, if the examinee returns to the exam following the above described scenario and changes numerous answers from wrong to right, or speeds through the balance of the exam and obtains an unusually high percentage of correct answers in comparison to the first part of the exam, the test sponsor then has collateral evidence that, in combination with the statistics paints a clear picture that the examinee engaged in intentional misconduct for the purpose of achieving an unearned advantage on the exam.

The previous hypothetical scenario is only one of many scenarios that can develop in the middle of an examination where a brief interruption of the examination for

purposes of conducting a limited investigation can prove invaluable. Furthermore, this example illustrates the necessity of collecting and preserving collateral evidence to add value to statistical evidence that, while meaningful in its own right, does not provide the test sponsor with evidence regarding the ethics, character or professionalism of the examinee.

After the Examination

Gathering meaningful collateral evidence of test fraud after an exam administration is already completed is one of the most challenging aspects of any test fraud investigation. It is for this reason that test sponsors would be wise to implement robust measures to prevent and detect cheating prior to and during the exam. Indeed, the collateral evidence collected and retained by the test sponsor prior to and during the exam can be immensely helpful to the investigator who begins a test security investigation after the exam is already completed.

Hypothetical Post-Exam Investigation Prompted by Extreme Similarity Data

If a statistical analysis of exam results suggests the likelihood of cheating based on the extreme similarity of responses among a group of examinees, the test sponsor should be able to quickly determine from its database whether any of the examinees in the group share one or more of the following characteristics:

- physical address;
- place of birth;
- email address;
- static IP address;
- payment information;
- educational institution; and/or
- employer.

Additional collateral evidence that the test sponsor should gather under this hypothetical scenario includes a complete review and analysis of all social networking sites of which the examinees are members, including but not limited to Facebook, Twitter and LinkedIn. The value of social networks as an investigative tool cannot be understated in the context of a test security investigation. To the extent an examinee shares any part of his or her social networking profiles publicly, the test sponsor can determine whether any of the people to whom the examinee is linked also took the exam and, if so, analyze that person's exam results for further evidence of collusion.

In addition, the test sponsor investigating test fraud will want to view the video recording and listen to the audio recording of the exam, check-in and the locker areas of the test center to see if the proctor might have missed any unusual behavior during the exam. The break schedule is another important piece of evidence because of the examinee's response conduct following breaks. Did the examinee return to the exam following a long break and change a series of answers from wrong to right? Did the examinee quickly answer some of the most difficult questions after taking a longer amount of time before the break to answer easier questions? Computer-based testing analytics and other forensic test data analyzed in conjunction with collateral evidence

gathered before, during and after the exam can prove instrumental in gathering evidence related to these questions and in guiding the test sponsor to reach reliable conclusions about the examinee's conduct.

CONDUCTING INTERVIEWS

Another important and productive method of gathering evidence following the exam is conducting interviews. The test sponsor's investigator should certainly interview proctors and any other person who observed the examinee's unusual behavior during the exam, or has other relevant information related to the investigation. However, of paramount importance in a test fraud investigation is the interview of the examinee suspected of possible involvement in cheating. The interview of an examinee suspected of test fraud should usually be one of the last steps in the investigation, so that the test sponsor has the opportunity to conduct the interview with the benefit of all available evidence and to ultimately confront the examinee and request that he or she explain the evidence. Nevertheless, investigators must always treat the examinee respectfully and make clear that the purpose of the interview is to gather facts and evidence, not to make accusations or judgments.

All examinee interviews should be conducted in person, so that the investigator can evaluate the credibility of the examinee based upon nonverbal cues and body language. In addition, a second person should always be present as a witness for the interview. Although a recording device can be used in place of a witness (with the consent of the examinee), recording devices may reduce the examinee's comfort level, chill the discussion and could create legal issues because the recorded statement of the examinee would have to be turned over in discovery in any subsequent lawsuit that relates to the matter under investigation. Depending on what the examinee says during the interview, the test sponsor may or may not want the verbatim recording of it made part of discovery in litigation. The problem is that the test sponsor would not be able to make that evaluation until well after the recording was made. That is why the author of this chapter generally advises clients to use two interviewers to conduct initial investigative interviews rather than using a recording device. Following the initial interview and after the test sponsor has an understanding of what the examinee will say, the test sponsor may then decide to take a recorded statement from the examinee.

When interviewing the examinee, investigators should remind the examinee of his or her obligations under the examinee agreement, including (hopefully) the requirement that the examinee cooperate in the test sponsor's investigation. Investigators should ask a series of background questions to which the answers are already known to the test sponsor, to establish a baseline for the examinee's veracity. During the interview, the investigator should ask questions that relate to all relevant evidence obtained prior to the interview.

For example, continuing with the hypothetical example of an investigation arising from a finding of extreme similarity of responses among a group of examinees, the investigator should ask about how the examinee prepared for the exam, including identifying all prep courses and prep materials used to study for the exam, and the sources of those materials. The investigator should ask about the examinee's colleagues, classmates, family members and friends to determine whether any of them either previously took the exam or took it at the same time or in close proximity to his or her exam. If any close associates of the examinee took the exam, the investigator should ask about how each of them prepared for the exam. The investigator should also ask in

detail about the events on the day of the exam, including asking about what he or she did before and after the exam, and whom he or she saw and spoke to throughout the entire day. There may be many other areas of inquiry that arise during the interview, depending on the nature of the existing evidence.

By the conclusion of the interview, the examinee should be asked to explain each piece of evidence that could indicate that he or she engaged in intentional conduct to obtain an unearned advantage on the exam. Finally, the investigator should ask the examinee to produce relevant documents and other evidence that relate to the investigation and the issues discussed during the interview. For example, if the examinee has a record of poor academic achievement in college, but obtained a very high score on his or her first attempt at the exam in question, the examinee should be asked to explain that. Consider, for example, an examinee who reports that he or she had a parent suffering from a terminal illness throughout the examinee's college career and that he or she was the only family member caring for the parent during those years; in such a situation, the investigator should respectfully request documentary evidence to support that.

Furthermore, any mention by the examinee of email or text communications would require the investigator to follow up by requesting evidence of those communications. If, for example, the examinee tells the interviewer that he or she participated in a study group and that a member of the group circulated an email with a study outline attached, the investigator should have the examinee access the email account during or at the conclusion of the interview, in the presence of the investigator, and forward the relevant emails and attachments to the investigator. Indeed, if there is any indication of multiple emails relating to examination preparation, the investigator should inspect and search the email account and direct the examinee to forward each and every email that the investigator deems relevant. If the examinee says that he or she sent her friend a text message during an exam break with a question about the exam, the message would have to be documented from the examinee's mobile device by the investigator examining the device and taking photographs of the messages and the contacts to preserve the evidence. If the examinee said that all of the members of his or her study group obtained exam prep materials from the same test prep company, the investigator should obtain a copy of the study materials and further pursue all investigative leads related to the company.

All study materials and documents obtained from the investigation that could potentially contain exam content must be analyzed by the test sponsor to determine whether there are any matches to actual exam content. Some test sponsors have software that performs these comparisons, but others may simply rely on test development staff members to manually make these comparisons.

REACHING CONCLUSIONS AND DETERMINING NEXT STEPS

Following the conclusion of the investigation, the test sponsor must evaluate all of the evidence it has gleaned from the investigation and reach conclusions about the validity of the examinee's score and whether the examinee engaged in conduct that falls short of the ethics, character and/or professionalism standards required by the test sponsor. At the conclusion of most investigations, it is rarely one piece of evidence that will dictate the test sponsor's findings. Rather, it is a collection of evidence in the aggregate that necessarily guides the test sponsor to particular findings.

The test sponsor must have clear evidentiary standards by which it can weigh the evidence and decide what actions it may take with respect to an examinee suspected of test fraud. The evidentiary standards applied by the test sponsor should be clearly spelled out in the examinee agreement, along with all potential consequences for findings of score invalidity and/or exam misconduct.

The highest evidentiary standard under United States law is that used for criminal cases: beyond a reasonable doubt. Although somewhat difficult to define, a reasonable doubt is generally interpreted to mean that a reasonable person would hesitate to find that the allegations are true. If a reasonable person would not hesitate in any way to act based on the evidence presented, then the matter is proven beyond a reasonable doubt. The much more lenient standard for a finding of civil liability under U.S. law is a preponderance of the evidence. The preponderance standard is best defined as a finding that it is more likely than not that an event occurred as alleged.

Test sponsors concerned with the ethics, good character or professionalism of their examinees may reasonably conclude that if a preponderance of the evidence demonstrates that the examinee engaged in conduct designed to obtain an unearned advantage on the exam, that is sufficient not only to cancel the examinee's score but also to impose sanctions in accordance with the sponsor's policies and procedures. Indeed, depending on the extent of the misconduct, some test sponsors may consider the examinee's misconduct a disqualifying circumstance that precludes the examinee from obtaining the credential sought through the test sponsor. On the other hand, some test sponsors may decide that the evidentiary bar should be beyond a reasonable doubt before they impose sanctions on an examinee that could have career ending or lifelong collateral consequences.

If the test sponsor concludes that the evidence of examinee misconduct fails to meet the evidentiary standard that it has adopted, but the examinee's test score is nevertheless not valid based upon a statistical analysis, the test sponsor can simply cancel the examinee's score and require the examinee to retake the exam without imposing any additional consequences (assuming that these consequences are clearly articulated in the examinee agreement).

CONCLUSIONS

Although statistical analyses are an important tool to detect testing irregularities and potential test fraud, it is useful to supplement scoring data produced by the administration of examinations to understand the events that produced those data. Test sponsors can gain a much better understanding of the events that cause aberrant scoring data by collecting and analyzing collateral evidence as described in this chapter. As demonstrated by the hypothetical scenarios presented in this chapter, collateral evidence gathered before, during and after the administration of an examination can provide crucial information that helps explain aberrant test results and enables the test sponsor to determine whether examinee misconduct caused aberrant exam results.

Although not all test sponsors are concerned with examinee misconduct, test sponsors that establish and maintain standards for the ethics, character or professionalism of their stakeholders have a strong interest in understanding whether an examinee engaged in conduct that demonstrates an examinee's failure to meet those standards. Yet scoring data alone does not enable test sponsors to draw any conclusions about any examinee's conduct. Thus, collecting and analyzing collateral evidence that helps explain aberrant test results also provides test sponsors with a unique opportunity to

assess the ethics, character and professionalism of its examinees. Only after a test sponsor is able to determine an examinee's conduct with a reasonable degree of certainty can it then make an informed decision about what actions it should take, if any, in relation to the examinee's score, eligibility, credential, certification, or license.

NOTES

1. No part of this chapter or any related presentation by the author constitutes legal advice, and no attorney-client relationship is created between the reader and the author.
2. For purposes of this chapter, a test security incident is defined as any event or set of facts that could potentially affect the validity of examination results. Examples of test security incidents include, but are not limited to, harvesting and/or copying of secure examination content by any person with access to the exam; administration of an examination to an imposter instead of the duly registered examinee; violations of policies and procedures during examination administration; and conduct by any person that could provide an examinee with an unearned advantage on the exam—conduct falling into this last category is commonly referred to as cheating.
3. Although not the subject of this chapter, test sponsors must have privacy policies that govern their collection, preservation, protection and use of personal information from examinees. Test sponsor privacy policies and practices must comply with a myriad of state and federal laws in the United States, as well as laws that apply in foreign countries where test sponsors may offer examinations. Test sponsors should be aware that there are laws that limit the collection and use of certain categories of information, including, for example, educational, financial and health information. Some laws also limit the categories of people from whom organizations can solicit information online, including, for example, children under the age of thirteen. Test sponsors should consult with an attorney to ensure that they understand all applicable United States and international laws, and establish policies and practices that comply with them in all respects.

REFERENCE

American Board of Medical Specialties. (2012). *ABMS statement on examination security*. Available from www.abms.org/News_and_Events/Media_Newsroom/features/feature_ABMSStatementOnExamSecurity_01132012.aspx



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Section IV

Conclusions



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

20

WHAT HAVE WE LEARNED?

Lorin Mueller, Yu Zhang, and Steve Ferrara

INTRODUCTION

This chapter is not meant as a comprehensive review of the relative strengths and weaknesses of each of the methods and concepts covered in this book, but as a high-level overview and practical perspective, taking into account the common themes across its chapters. The chapters in this book provide a comprehensive presentation of techniques, both statistical and practical, for detecting and dealing with cheating in its various forms. As a first step, we will cover the basic approaches, findings, and interesting points made in each chapter. Next, we will compare the results for one of the example databases to determine how well these approaches do as a whole in detecting cheating. Next, we will discuss some broad themes from across the chapters, noting chapters where an exceptionally informative discussion of a particular theme can be found. Finally, we lay out five challenges for researchers and practitioners using these methods to detect cheating on tests.

OVERVIEW OF METHODS COVERED

Zopluoglu, in Chapter 2, begins this volume with an examination of the status quo with respect to multiple response similarity measures and person-fit measures. Results indicated that the response similarity indices are superior to the person fit indices as measured by area under the receiver-operator characteristic curve (AUC), which is to be expected given the nature of the cheating in the licensure dataset (a selection of items harvested and shared). We agree that AUC is the best method for evaluating the performance of these indices under various conditions, but it is possible that person-fit indices would outperform similarity indices under a specific flagging policy. Readers are advised to inspect the receiver-operating characteristic curves to determine whether there are specific flagging policies that would contradict the general finding.

In the third chapter, Maynes provides an overview of the rationale behind similarity analyses, focusing primarily on *M4*, and provides several exhibits to explain deviation from expectations to stakeholders. This chapter raises the issue that other factors can explain similarity, such as instructional bias or studying in groups, answering strategies,

and form reuse. These are framed as violations of assumptions, but to the extent that they are likely to occur, they are a problem for interpreting these analyses. However, the excellent graphical representations of similarity can be helpful to see whether similar test instances are a matter of a few items or an extremely suspicious outlier.

Kim, Woo, and Dickison follow in Chapter 4 to outline a selection of person-fit and a machine-based learning approach (market basket analysis) to detecting aberrant response patterns. CUSUM in particular is an intuitive way to explain aberrant response patterns to stakeholders with little psychometric expertise, and often long response strings that don't fit within expected patterns can be presented in a way that is apparent to a casual observer. Machine-based learning approaches remain a mixed bag. Often, they can be helpful to determine risk factors for cheating, but other times they can produce findings that are not helpful. These approaches are heavily dependent on the data made available to the machine learning algorithm, and the appropriateness of the technique to exploring the data.

Thinking about these three chapters together, it is important to note that similarity and copying indices must be interpreted carefully. Highly similar answers might not be aberrant (i.e., they might represent plausible response patterns), and aberrant answer patterns can happen even when examinees don't have any preknowledge or are unable to copy from one another. Similarity can be the result of predictable (but not compromised) test content, studying strategies (such as studying heavily weighted content areas exclusively), common responding strategies (such as "always guess C"), or failing to account for administrative issues such as test speededness. Similarity indices are best interpreted alongside information about item compromise.

With that in mind, Eckerly, in Chapter 5, tackles the challenge describing the status quo for identifying item preknowledge and compromise covering a variety of methods, including the Deterministic Gated Item Response Model (DGM), moving averages, log odds ratio, and a combination of Differential Item Functioning (DIF) and Differential Person Functioning (DPF). Eckerly makes several helpful points about the conditions under which each method is appropriate, and how factors such as the proportion of compromised items affects the performance of each statistic. The chapter provides an excellent set of recommendations for practitioners to help them better understand how these indices will perform and should be interpreted in their testing programs.

In Chapter 6, Wollack and Maynes focus on *M*₄, which is perhaps the most widely used and researched similarity index in practical settings. As expected, *M*₄ performed well except in cases where cluster sizes were small. The authors make a good point that similarity clusters might be detected most effectively by using more stringent criteria to detect clusters of examinees that show extreme similarity and then building up a larger cluster by identifying other examinees similar to them.

In Chapter 7, O'Leary and Smith use DIF and DPF to detect both items and persons that do not fit the response model. Although this approach had only moderate success at detecting item compromise and known cheaters, it is intuitively appealing because it uses methods that are familiar to many professionals in the testing industry. Another interesting finding is that some items are flagged in the opposite of the predicted direction, which could indicate that a substantial proportion of the test content is compromised or predictable, making some uncompromised or unpredictable items appear to be more difficult than expected by the (now biased) response model.

In Chapter 8, Belov makes the point that preknowledge is not a monolithic entity; different examinees can have access to different sets of compromised items. Belov used a divergence algorithm to simultaneously detect groups of compromised item subsets

and subgroups of examinees with preknowledge of one or more of those item subsets, essentially looking for large performance differences on groups of items that can be linked to clusters of examinees. This technique is probably best considered an exploratory technique in its current state. It performed very well in simulation data, but had a high Type I error rate in the real data (although the author accurately notes that there might have been cheaters that were not identified by the test sponsor).

In Chapter 9, Boughton, Smith, and Ren examined using response time patterns to detect compromised items and cheaters on computer-based tests. The method was successful at detecting some of the items designated as compromised in the licensure dataset, but for many compromised items there was no strong response time pattern. It would be interesting to look at the properties of detected items as opposed to undetected items (length, use of graphic stimuli, tables, etc.). This model might be much more successful in situations where response time difference would be more apparent, or where items are harvested with a very high degree of veracity.

In Chapter 10, Bishop and Egan turn our attention to the status quo in erasure analysis, using multiple approaches ranging from the fairly simple Z score methods, to more sophisticated considerations of the joint distributions of erasure types, HLM, and generalized estimation equations (GEE). They also present some compelling visualizations to emphasize practical significance. The drawback to these approaches is that they are limited in their ability to identify cheating individuals when cheating is not occurring within an already defined group (such as a classroom or test site), and did not have high agreement levels among the different approaches (perhaps further arguing for looking more closely at measures of practical significance alongside inferential tests).

In Chapter 11, Wollack and Eckerly continue on the theme of erasure analysis, extending the use of the erasure detection index (*EDI*; a variant of the omega statistic) to group-level cheating. Authors note precautions taken to account for valid situations of many wrong-to-right (WR) erasures (e.g., skipping or double-bubbling a question). Authors note that group-level (class, school, district) *EDI* did not correlate well to practical measures of cheating (such as the number or percentage of WR answers), but we might expect to see much larger correlations in a scenario where there is more substantial cheating.

In Chapter 12, Skorupski, Fitzpatrick, and Egan use a Bayesian hierarchical linear modeling approach to detect cheating groups. Consistent with expectations for an HLM model the within group variance had a noticeable effect on power, as did the Posterior Probability of Cheating, which would be expected for a Bayesian model. These results are both promising and troubling: promising in the sense that if we have good, long-term data that we can use as baselines to examine these factors we can get accurate results; troubling in that they could give organized cheaters a countermeasure to fraud detection (i.e., by inflating variance through encouraging low performance among a few examinees).

In Chapter 13, Clark, Skorupski, and Murphy used cumulative logit regression to identify classrooms with a greater than expected proportion of students in a performance level based on their score from the previous year. As was the case with several chapters, the model was limited by the lack of additional explanatory variables; however, including those variables can cause issues related to patterns of data missing not at random (MNAR; i.e., not imputable or ignorable by virtue of data that are included in the model). Although multiple imputation (MI; Rubin, 1987) was suggested by the authors as a solution to the missing data problem, one must wonder how such a process would be viewed by stakeholders without expertise in statistical prediction. As

recommended in Chapter 6, we might think about using such a technique to identify cases for which we want to do additional investigative work.

In Chapter 14, Gaertner and McBride found the two-proportion Z score to be a better indicator of cheating at the group level (collusion or erasures) than a more sophisticated multilevel logistic regression (MLR) model for detecting pass rate changes at the group level across testing years. However, the authors admittedly hamstrung MLR by not including covariates so that it could be more directly compared with the Z score method. Using available covariates could allow the MLR to outperform the simpler model, a theme raised in Chapter 13 and echoed throughout other chapters. Even using the simple Z score method, the authors note that other predictors can be used to help interpret and strengthen the statistical findings.

In Chapter 15, Martineau, Jurich, Hauger, and Huff provide a good overview of issues and recommendations for testing programs as they further invest in up-to-date technology for their delivery platforms, particularly those coming out of the large collaborative efforts for state accountability systems. Their recommendations present and summarize expert views on handling complex data and investigations and are an important step toward promoting industry standards for forensic data use in a high-stakes testing environment.

In Chapter 16, Harris and Huang make the case for establishing reliable baseline data to serve as an interpretive frame for statistical tests designed to detect cheating. This recommendation was raised in several preceding chapters, and the authors do an excellent job of providing guidelines for collecting such data and presenting the information in graphs, graphics, and tables.

In Chapter 17, Foley presents a range of visual techniques that can help practitioners to explain the results of statistical tests designed to detect cheating. For the most part, these recommendations are easily achievable with basic spreadsheet software.

In Chapter 18, Skorupski and Wainer make a reasonable case for adopting a Bayesian approach to conducting analyses to detect cheating. The chapter briefly presents the rationale behind Bayes' Theorem, and arguments for why this approach is superior in many ways to the more common frequentist approaches. The lure of Bayesian analysis is strong, and as efforts such as this book help the industry to accumulate better information about the prevalence and risk factors associated with cheating, it becomes more reasonable to think we can generate more accurate priors. Practitioners who have not considered adopting a Bayesian perspective within their testing programs should consider this chapter carefully, and ask themselves whether they have enough information to take advantage of Bayesian techniques.

Chapter 19 serves as a fitting conclusion to the chapters covering technical and practical recommendations. In this chapter, Weinstein covers practical recommendations for examining biographical data gathered through registration and steps for conducting cheating investigations. The chapter includes recommendations for gathering additional evidence, conducting interviews, and reaching conclusions based on the evidence. Consistent with the recommendations outlined in other chapters, the author recommends developing clear processes and standards in advance of taking action during an investigation.

HOW DID WE DO?

Many methods have been proposed and applied in this book for detecting test results of cheating behaviors. Some chapters analyze the same data and make it possible to compare the consistency of detection against what we already know in the data.

The authors from six of the chapters in this book graciously submitted their findings from the credentialing dataset, as described in Chapter 1, for our review (the K-12 data were less amenable to this purpose, as the test sponsor provided no information on the hypothesized compromise status of the examinees or items). The analyses on this data share a common goal of detecting examinees identified as having preknowledge of some test content. Where possible, we collected results of the analyses from the chapter authors to compare them. Comparing flagged examinees across analyses can help us understand the strengths and limitations of each analysis in detecting cheating behaviors, specifically the case of having preknowledge of test content. We have chosen not to identify the authors here. We merely want to present some correspondence data to better our understanding of the cheating detection problem.

Hits, Misclassifications, and Specificity

Our first comparison focuses on true positive hit rates. A flagged examinee that matches the test sponsor's list of known cheaters is considered as a hit (true positive). The hit rate, sometimes called sensitivity, reflects the proportion of listed examinees flagged by an analysis. Among the six datasets submitted for our review, several of the author sets applied multiple methods or used a variety of cutoff values, and thereby generated more than one set of flagged examinees. For this reason, mean, minimum, and maximum of hit rates are reported.

Another measure of the effectiveness of a signal detector is specificity. Specificity is the proportion of true negatives (noncheaters) that are correctly identified as negatives—in other words, the proportion of cases indicated as “not cheating” by the test sponsor that were also identified as “not cheating” by the detection mechanism. For a full discussion of these and other measures of signal detection, including strategic considerations for when particular measures should be maximized, see Mueller and Munson (2015).

Sensitivity and specificity tell how well an analysis does in detecting examinees who might have access to test content and those who do not. In many practical cases, we are also interested in the combined rate at which we misclassify examinees. For this analysis, we calculated the total number of examinees who were classified as cheaters but were not listed by the test sponsor (false positives) and the total number of examinees who were listed by the test sponsor but not detected by the algorithm (false negatives), as a proportion of total examinees. This rate represents the combined rate of Type I and Type II errors.

Table 20.1 presents the hit rates (sensitivities), specificities, and misclassification rates for the six sets of submitted findings. Hit rates ranged from 0.0% (Author Set C, Author

Table 20.1 Classification Rates (%)

Analysis	Hit Rates/Sensitivities			Specificity			Misclassification Rates		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
Author Set A	11.1	8.3	14.6	98.7	98.6	98.8	4.0	3.8	4.1
Author Set B	28.6	14.6	41.7	87.7	79.9	94.5	12.1	4.8	13.4
Author Set C	0.0	0.0	0.0	97.9	97.9	97.9	2.1	2.1	2.1
Author Set D	15.6	0.0	50.0	91.7	67.1	99.9	8.1	0.1	31.9
Author Set E	19.8	12.5	27.1	99.4	99.2	99.7	6.6	2.9	10.3
Author Set F	8.3	8.3	8.3	100	100	100	2.6	2.6	2.6

Set D) to 50.0% (Author Set D). Specificities were high, ranging from 67.1% (Author Set D) to 100% (Author Set F). Misclassification rates ranged from 2.1% (Author Set C) to 31.9% (Author Set D), and correlated with the hit rate. That is, Author Sets B and D had the highest hit rates, but also the highest misclassification rates. Taken together, the results in Table 20.1 demonstrate that a great deal of the control over who is flagged lies in the choice of cutoff values. Doubtlessly, these values change substantially given the appropriateness of the detection method for the cheating methods in use, and the nature of the examination data.

Agreement Rates

Another important step in the validation of these analyses is to identify examinees on which they reach the same conclusion. We looked at this in three ways: (1) the joint flagging rate, or the rate at which both methods flagged the same examinee (irrespective of whether that examinee was flagged by the test sponsor); (2) joint hit rates, or the rate at which both methods flagged an examinee identified as a cheater by the test sponsor; and (3) joint misclassifications, or the rate at which both methods flagged an examinee not identified by the test sponsor. The results of these analyses are presented in Table 20.2.

Overall percentage agreement in Table 20.2 is low. In many contexts, these would be disappointing findings. However, these analyses are each attuned to specific forms of cheating, so it is difficult to tell whether they are insensitive to cheating in general or merely poorly suited toward the forms of cheating in the data provided. Author Set C failed to identify any instances of cheating identified by the test sponsor, and had low joint flagging rates with the approaches, suggesting this analysis was especially not well-suited to the data. Author Set D had the highest agreement with the other authors, but also produced the highest misclassification rates.

We also looked at what would happen if we combined all the detector results into a single decision and flagged any candidate that was flagged by any of the detectors used by any of the authors who submitted findings for review. Using this method, the resulting hit rate was 62.5%, the combined specificity was 42.7%, and the combined misclassification rate was 56.8%. In other words, although we were able to detect the majority of cheaters listed by the test sponsor, we did so by generating so many false positives that the sensitivity value dropped substantially and the misclassification rate rose by multiples.

Table 20.2 Agreement Rates (%)

Analysis	Joint Flagging Rates					Joint Hit Rates					Joint Misclassification Rates				
	B	C	D	E	F	B	C	D	E	F	B	C	D	E	F
Author Set A	1.1	0.0	1.0	0.7	0.1	10.4	0.0	10.4	10.4	4.2	2.5	2.6	1.9	2.6	2.5
Author Set B	—	0.2	13.2	3.1	0.2	—	0.0	39.6	25.0	8.3	—	1.9	13.2	4.0	1.7
Author Set C	—	—	0.9	0.3	0.0	—	—	0.0	0.0	0.0	—	—	1.8	2.4	2.7
Author Set D	—	—	—	4.5	0.2	—	—	—	22.9	6.3	—	—	—	4.9	1.2
Author Set E	—	—	—	—	0.2	—	—	—	—	1.2	—	—	—	—	2.1

Limitations of the Comparison

Although these numbers present a somewhat bleak view of using statistical analyses to detect cheating, a few limitations must be noted. First and foremost, the authors were working with a limited amount of data they were unfamiliar with, which is undoubtedly a hindrance to fully understanding and analyzing the data. Expertise in dealing with complex data from a particular testing program can take years to develop, and these authors had no opportunity to ask for additional helpful information that might exist in the test sponsor's databases. As such, these results represent what would be a first attempt at detecting cheating. Taken in that light, the high hit rate is much more promising than if we were considering the final results of a complete investigation.

Second, as mentioned previously, the authors had to select a cutoff with little regard for the relative consequences of their decisions. Are they selecting cutoff values to invalidate scores without an investigation? Are the flags going to be used to begin an investigation? If so, what resources are available to the organization to carry out these investigations? Does the test sponsor want to contact anyone whose test data warrant any suspicion of cheating, or do they want to focus their energies on only the most egregious cases? Here again, decisions about cutoff values for cheating detection analyses are best, hopefully typically, selected with an understanding of organizational resources, likely forms of cheating and realistic prevalence of each, and the consequences of Type I and Type II errors in statistical flagging. In practice, we would hope that cutoff values for various analyses are set with more consistency than is possible in this exercise.

Third, the submitting authors did not have extensive background on the pervasiveness or specific mechanisms for cheating (e.g., erasures, copying, item harvesting). Some information was provided, but additional information might have been extremely helpful in determining which methods would be best suited to detecting cheating, as well as what proportion of candidates would be ideal for maximizing the hit rate while minimizing false positives. As mentioned previously, the true extent of cheating was unknown to the test sponsor, so even some of the false positives identified by these analyses might have been cheating.

In all, these methods in isolation did not do an exceptional job at detecting cheating, as identified by the test sponsor. However, in combination, they were able to detect the majority of cases of cheating identified by the test sponsor. False positives remained a significant problem, but we don't know, and most programs will never know, the extent to which flagged examinations that can't be corroborated as instances of cheating are false positives, or whether they are investigative "misses." It is also important to note several factors when considering these findings. As some of the chapter authors note, there are likely other instances of cheating in the provided data that were not able to be confirmed by the test sponsor. As such, the false positive rate might represent a slight overestimate of the true value (but a realistic value when considering which instances can be acted upon). Second, the hits were themselves initially identified, for the most part, through forensic analyses, and confirmed with additional evidence. Given the role of forensic analyses in most testing programs, that being to initiate investigations and direct resources to the most appropriate targets, the results can be considered very promising in practical context. That is, if we were acting in a vacuum without the possibility of additional information to confirm or reject the initial findings, we would want better classification performance. But given our ability to collect additional information, or in some cases merely consider additional information already available, lower performing classification decisions are entirely acceptable.

COMMON THEMES

In many ways, what we learned was not surprising: statistics used to detect cheating and other illicit or unsanctioned examinee behaviors function like many other statistics. They are useful to the extent that the base rate lends itself to detection, as most directly shown by the Skorupski and Wainer chapter. For good or bad, very low base rates exacerbate the false positive problem with many of these statistics. In addition, we need to make sure our data are clean and that we have selected an appropriate standard for comparison, as demonstrated in the Harris and Huang chapter. Poorly chosen baseline data can inflate false positives or false negatives, or both, depending on the nature of the error. Finally, in most testing environments, we must consider a variety of examinee and group characteristics in selecting a standard for comparison: testing history, testing accommodations, comparability of the opportunity to learn the material, and language proficiency, to name a few. Following the many practical and scientific recommendations raised throughout this book may not only help to reduce the problem of false positives but also increase cheating detection rates, improve the certainty of our conclusions, and improve the efficiency of investigative processes.

In looking across the chapters, it should become clear to the reader that each approach is tailored to a particular set of circumstances and assumptions. It is worth considering the basics of your testing program, including the administration procedures, highest priority security threats (Ferrara, 2016), and the nature of the construct being assessed before moving forward with any particular analytic strategy. For pencil-and-paper testing, where the security of examination materials is constantly a concern, erasure analysis and similarity indices that focus on copying may be the most appropriate. For tests with long testing windows or high rates of form reuse, item parameter drift, fit analyses, and similarity indices that focus on possible item harvesting may be more appropriate. For large-scale computer-adaptive tests or tests with highly linear construct progressions (e.g., mathematical reasoning), fit analyses might be the most useful, whereas for diverse constructs such as job knowledge, fit might not be as useful.

Little attention is paid to norm-referenced as compared to criterion-referenced tests, but this aspect of a testing program is critical in determining which examinees or groups to investigate and sanction. For norm-referenced tests, an advantage of a few items (whether gained by preknowledge, copying, erasing, or other means) can have a practical impact of moving an examinee ahead of hundreds or even thousands of other examinees, or result in a large-scale score change at the extremes of the scale. For criterion-referenced tests, having an advantage of a few items only matters when it moves the candidate from one classification into a higher classification. In some cases, the nature of the test score can change across levels, such as when a criterion-referenced test in K-12 testing gets aggregated into a norm-referenced comparison at the classroom or school level. Taking this basic information about the test uses and consequences into account can help testing organizations to make better-informed decisions about the trade-offs between detection power at various places in the score scale and false positive rates. Of course, any cheating affects the integrity of the scores produced, but we should be especially concerned about those that have the greatest impact on the validity of the interpretation of those scores.

Looking beyond which indices might be most useful in a particular setting, the chapters in this book should serve as strong evidence that relying on any single method is unwise. Using the chapters in this book, it would be a worthwhile task to create a matrix of all possible mechanisms for cheating that relate to a testing program with

all possible exam security analytic methods, and match which methods best relate to the cheating mechanisms relevant to that program. Moreover, developing an understanding, over time, of how indicators relate to one another, within a testing program, is important to developing more complex indicators of cheating. We will discuss this point in more depth later in this chapter.

Ferrara (Ferrara, 2014, 2016) provides a guide to prioritizing threats to test security and deciding which require attention. The guide is part of a larger framework for building comprehensive test security systems: Prevention, Detection, Investigation, and Resolution. As is clear in Table 20.3 (columns 2 and 3), postadministration forensic analysis addresses a limited number of cheating and other threats to test security.

We have also learned that gathering extensive data, outside of what is recorded during the administration itself, can be extremely helpful in gaining a better understanding of the plausibility of an exam score. These data can include previous testing history, biographical information, home and work addresses, employers, educational

Table 20.3 Countermeasures for Cheating and Other Threats to Test Security

Before Test Administration	During Test Administration	After Test Administration
Examinees		
Acquiring test items: Management of chain of command of secure printed materials and protection of access to digital files	Copying or supplying answers, using cheat sheets: Management of materials brought into testing sessions, observation of examinee behavior, intervention on suspicious behavior	Divulging secure test material: Postadministration surveys; observations and surveys about test preparation activities prior to the next administration
Test Administrators		
Divulging/teaching secure material: Postadministration surveys	Providing answers, indicating answers that should be changed: Response similarity analyses; test administration observations	Changing answers on answer documents: Erasure/wrong-to-right answer change analyses
Other Staff in Local Testing Sites		
Failing to publicize expectations for professional behavior or provide training; failing to monitor management of chain of command: Communication and training	Failing to monitor test administrations and test content security: Systematic monitoring of administrations	Failing to acknowledge, report, investigate, or resolve violations: Policies and practices to support enforcement and follow-through
Testing Program Managers and Operations Contractors		
(a) Failing to provide effective test administration and security training: Comprehensive training plans and policies to ensure all test administrators and local testing site staff participate	Note: Here, testing program managers and vendors have to rely on local testing-site staff	(a) Failing to account for all secure test content and chain of custody and protection: Systematic accounting for all material and documentation and enforcement of custody chains
(b) Failing to effectively engage in efforts to discover cheating: Web monitoring and promoting simple channels for anonymously reporting security breaches (e.g., tip lines, physical mail or email security inboxes)		(b) Long testing windows and test forms reuse: Item fit, item parameter drift, and response similarity analyses

background, preparation activities, and grouping variables, such as school or classroom. Each of these pieces of information can be used to better inform models of the plausibility of examinee responses and test scores, or to better identify information sharing networks among examinees. More reliable information can drive investigation resources and help to guide decisions about which forensic analyses are best suited to your program and decisions about how to best implement those analyses. These decisions include better narrowing groups for computationally intensive similarity analyses, identifying when score gains are explainable, or better estimating the likelihood that an examinee gets a difficult item right. Exploring data through social network analysis techniques could be extremely useful in detecting formal and informal networks of cheating behaviors (Brass, Butterfield, & Skaggs, 1998). Network analysis can be used to detect patterns of similar home addresses (e.g., examinees living together or nearby sharing recalled items), educational institutions frequently associated with suspicious test scores (e.g., possibly maintaining a database of test content), or preparation activities (e.g., test prep providers harvesting items).

These chapters also illustrate that some tried-and-true statistical techniques can be applied to test security matters, most notably in the O'Leary and Smith chapter examining Differential Item Functioning (DIF) approaches to detect instances of preknowledge. The Gaertner and McBride chapter on comparing pass rates and the Clark, Skorupski, and Murphy chapter on nonlinear regression also explore relatively well-known statistical techniques employed in a test security setting. The Gaertner and McBride chapter has an excellent discussion of comparing techniques in terms of power and false positive rates, providing further evidence that using known techniques give us some degree of certainty in the strengths and weaknesses of our methods. More thoroughly exploring these known techniques can help us build comprehensive test fraud detection systems from a solid base.

Several chapters mention the role of test security analyses in score *validation*. Any test security program must begin with validation as the central premise of everything it does. In fact, a testing program must be able to provide evidence to support the claim that test scores are validly interpretable because cheating and other validity threats (e.g., content bias, administration problems) have been eliminated (to the extent possible) as explanations for score variance in statistical analyses. We want to ensure that inferences based on our test scores can be relied upon to guide our actions. Even in the case where we cannot determine that an examinee has actively cheated, many of these analyses can be used to support an argument for test-score invalidation, especially when the analyses involve very high score gains or high degrees of misfit. A reasonable argument could be made, in many circumstances, that a particular score was not indicative of that examinee's true ability. Here again, test sponsors might want to conduct some organized thinking about whether the analyses they employ and the constructs they measure lend themselves to this kind of action.

SCOPE OF THE PROBLEM

It's hard to imagine any test sponsor reading these chapters and failing to recognize the enormous amount of resources one would need even to approach a comprehensive implementation of the techniques covered. One is reminded of the parable of the blind men feeling the elephant. In each of these chapters, we found evidence that the technique in question was able to identify a significant proportion of cheaters, but typically at the cost of a large number of false positives. The low detection fidelity is not a

problem of the techniques themselves or the abilities of the researchers. We have many capable researchers, each working on a distinct and important part of the problem, but in doing so we lack the integration process we need to improve detection. We're learning a great deal, but efforts like this book are the first essential step to moving away from this analogy and toward a more effective model of investigating this complex problem. A more promising analogy is that of Hercules slaying the Lernean Hydra. We must conceptualize test fraud as a problem that can manifest itself in many forms and be prepared to attack many heads at once. In other words, we cannot rely on a single approach (e.g., fit versus similarity) to detect cheating, and perhaps we should consider multiple statistical formulations within each approach. We also need an understanding of how indices relate. For example, fit indices by themselves might not indicate much in some testing programs, but when a high misfit value occurs simultaneously with unusual item response times and high similarity indices, we can build a strong case for score invalidation based on preknowledge. Using multiple indicators can help us to move away from artificial and sometimes arbitrary dichotomies based on a single indicator and toward a better understanding of the plausibility of a valid score based on multivariate data.

It's also notable that Hercules did not slay the hydra alone. Hercules relied on his trusted nephew, Iolaus, to seal each cut he made. Similarly, tackling the test fraud problem is impossible for any test sponsor acting alone. We need to work together as a profession and industry to establish standards and goals for record keeping, the qualifications and training of test proctors, and guidelines for effective investigations. We also need to encourage the leadership of our organizations, many of whom do not have the kind of background to quickly appreciate the meaning of these statistics, to be courageous in their pursuit of cheaters and defense of actions taken as a result of aberrant test data. Being too cautious in this area may engender a similar or greater level of risk than occasionally losing a legal battle. And like with the financial operations of our organizations, we must consider the importance of external consultants to supplement and improve our capabilities, and to auditors to review our processes and decisions. Again, Table 20.3 highlights the point that we have to rely on many different people in the testing process to prevent cheating and security problems and to detect and investigate them when we have reason to suspect cheating has occurred.

The most important benefit of thinking like Hercules is that we can move away from a model that relies on flagging on the basis of single indices and toward a process that identifies aberrant test data in a multivariate sense. Specifically, rather than focus on a single indicator of cheating, such as a similarity index, combining a "significant" amount of aberrance based on similarity indices with small but notable values of misfit, evidence that a candidate spent less time on items that showed misfit or were similar to other candidates' answers, or that their responses were similar to a cluster of candidates living or working in close proximity builds a better case for score invalidation than relying on evidence from a single analytic technique.

FIVE CHALLENGES

Thinking through the scenarios, statistical methods, findings, and issues described in these chapters raises a number of high-level challenges we must address to make test fraud detection analyses more actionable in practical settings. Although there are many challenges we could describe, we focus on five that come to the forefront for test sponsors and their technical staff as consumers of the information generated by these analyses.

Challenge 1: Dealing With Low Signal-to-Noise Ratio (SNR)

The main difficulty in detecting cheating is that unless cheating is pervasive and substantial, cheating occurs in an environment that we know is pervaded by many sources of variance (e.g., item sampling, human performance, administration issues, response transcription errors, misreading or misunderstanding items, predictable content). Even with tests of high reliability, the standard error of measurement might suggest that a score with many statistical flags is not that far out of the realm of possibility. In other words, test scores are noisy data, and it takes a substantial signal or pattern of smaller signals to be able to detect potential cheating within those data. Being able to cheat on a few items (the signal) might not be detectable among the nuisance factors that contribute to score variance (the noise). Thus the signal-to-noise ratio (SNR) is typically weak.

Challenge 2: Understanding Effect Sizes

Following on the issue of detecting small signals in noisy data, practitioners engaged in using statistical analyses to detect cheating need meaningful ways to discuss the impact of cheating on their testing programs. At the level of the examinee, knowing a few items in advance is helpful and can allow examinees to pass an exam if they are slightly below the cut score. One examinee misclassified due to cheating is a problem, but when that problem is repeated across dozens or hundreds of candidates, even the credibility of large testing programs can be undermined.

Across the chapters of this book we have seen a number of approaches to describing the extent to which cheating may have an impact on test scores. Some of them are readily interpretable by practitioners and laypersons, such as CUSUM (as described in the Kim, Woo, and Dickison chapter). Others are readily interpretable by statisticians, such as probabilities or indices that transform to Z scores. Still other indices are reported in values that can only be interpreted in the context of a specific examination with little similarity in cutoff values and interpretability across examinations of different lengths, item difficulty distributions, and candidate ability distributions. Some of these statistics can have vastly different expected distributions across different forms and testing occasions within the same testing program.

Beyond the difficulty of explaining these statistics to stakeholders who lack a background in psychometrics or statistics, even for psychometricians who work with these statistics regularly, it can be difficult to translate marginal differences into a metric that is inherently meaningful in terms of the impact on individual examinees, group performance metrics (where relevant), and the testing program as a whole. This theme has been repeated in several chapters. As we continue to research these methods, it stands to reason that at least some effort must be put into thinking about translating these indices into “effect sizes.”

When we see results from these kinds of statistical analyses translated into any sort of common metric, we often see them stated as “results this aberrant only occur by chance in 1 out of (insert unbelievably large number here) occurrences.” We will address whether this kind of statistical reasoning is even accurate later in this chapter, but even more troubling is that it tells us very little. First, it is unlikely that any particular test instance is completely independent of any other test occurrence (examinees study together, attend the same classes, use the same preparation materials). Second, it tells us very little, possibly nothing, in terms of how far the data differ from expectation.

Did this examinee get one more item correct, on average, than expected? Or did the candidate get ten more items correct than expected? How substantial is the deviation from the expectation? How substantial is the alleged cheating?

We need to think about how to transform abstract metrics into more tangible ones. That means reframing our results from eye-popping probabilities into more readily understandable metrics of test compromise. For some metrics, calculating effect sizes is relatively simple. For others, it could require computer simulation or establishing baseline data from which to compare results from tests with no detected cheating.

Moreover, considering basic elements of the validity argument of the test is helpful to determine how to best describe the results of these analyses. For example, many of the indices here focus on criterion-referenced tests and could frame the effect in terms of likelihood of passing without aberrant responses or the number of additional questions answered correctly. For norm-referenced tests, one might describe the change in percentile rank from possible cheating, or the number of candidates negatively affected by cheating.

Challenge 3: Making Sure Our Math Is Right, Even When We're Sure Our Math Is Right

Another common theme raised throughout the chapters in this book is the need for better information to inform our statistical models. To what extent are we accounting for information that we know is important to getting our models right, and to what extent do we ignore important information because it is hard to collect or difficult to account for in our models? Throughout these chapters, we have seen authors argue for more information regarding constructing proper clusters for pass-rate information, registration information, and selecting appropriate data for baseline comparisons. All of these additional data requirements suggest that our statistical models are not accounting for some very important pieces of information when it comes to determining whether an examinee has cheated or has validly achieved a score.

For example, if we consider responses to four-choice multiple-choice items, two candidates may appear to exhibit very similar response patterns in terms of the questions they get right and the responses to questions they get wrong. However, if we could account for such factors as the fact that these candidates were both educated in a language other than the testing language, and that they both relied heavily on one or two preparation guides to study for the examination, a moderate number of item correspondences that seem unlikely when compared to the typical examinee now seem more plausible. The statistics we used typically assume that observations are independent. In practice, this is rarely the case: we typically do see candidates of similar backgrounds accessing the same study materials and having common misunderstandings of test content based on curricular differences. That's not to say that these similarities can explain all of the commonalities in two examinees' response patterns, but they could explain some. If we can explain some of the correspondence in responses due to education, language issues, or studying patterns, we might be less concerned about a few items that are successfully harvested from a form.

A related problem with interpreting our results is our tendency to make cross-level inferences. So much of our test construction framework, and the statistical analyses described in this book, are built on the idea of a strong general ability factor. As a profession, we have a tendency to dismiss deviations from this general factor as measurement error, when it could be better described as unexplained or unmodeled variance.

As with the example above, educational experiences, study habits, and test-taking strategies can play a large part in test scores, and we need to be very careful about attributing seemingly systematic variance to cheating when we haven't collected adequate evidence to rule out the other possibilities. Even a very powerful general ability factor does not predict individual item responses very well.

We also need to adopt more consistent terminology. Models that accurately account for the way we use items within and across administrations may be more able to detect cheating than models that treat all items as if they have the same potential for compromise. Computer-adaptive and linear on-the-fly testing programs generally do well in describing item usage, but paper-and-pencil and nonadaptive computer based testing programs don't always think about item use in very thoughtful ways. In many cases, we see the terms *item use* and *item exposure* treated interchangeably, when various types of uses have very different implications for compromise. We propose the following definitions:

- Item uses: The number of times an item has been used on a separate form or form family within a testing window.
- Item views: The total number of examinees who have viewed an item.
- Item exposure: The extent to which an item has been potentially compromised. In some cases, item exposure will be equivalent to item views. In other cases, item views could be weighted by the likelihood that a particular candidate is harvesting items.
- Content exposure: The extent to which content has been compromised, as measured by the cumulative exposure of items with very similar content.

For example, consider the use of three test forms. Form A is administered to 20 examinees across two testing windows. Form B is administered to 20,000 examinees in one week-long testing window. Form C is administered to 10 examinees on one day, who are affiliated with known item harvesting groups. In many programs, Form A would be considered the "most exposed" because the content was used in two testing windows. However, Form B was administered to 100 times more examinees, and the content is much more likely to have been compromised. Form C is almost assuredly compromised, despite it being administered to the smallest number of candidates and the shortest testing window.

The Skorupski and Wainer chapter lays out the final issue with making sure our math is right: our math is frequently wrong, or at least interpreted incorrectly. As mentioned previously, many of the statistical techniques covered in this book describe results in terms of the frequency of occurrence given various assumptions of independence. Beyond the issues raised in the Skorupski and Wainer chapter, two issues undermine the usefulness of these statistics. First, they rarely include the appropriate interpretational context. Specifically, a dataset containing 1,000 examinees can generate approximately half a million one-to-one comparisons. That makes a 1 out of 1,000,000 probability seem much less persuasive. Second, stakeholders don't always have a great deal of confidence in hard-to-fathom numbers. It should not be surprising that probabilities that are hard to conceptualize are treated with skepticism, especially when they are associated with multiple exam records.

Challenge 4: Explaining These Results to Laypersons

Much of this volume has been devoted to convincing ourselves that we can identify cheaters with confidence, and for good reason. We must build the case to justify

invalidating a score or opening an investigation. We have mentioned several times in this chapter the difficulty of explaining the results of these analyses to stakeholders without a background in psychometrics or statistics. We won't repeat those statements here, but we will note that this is a challenge worthy of consideration by itself. Specifically, we need effective ways of communicating complex results to several stakeholder groups with differing interests, including

- test sponsor management and boards of directors, who are concerned about the financial, fiduciary, and public relations problems caused by cheating;
- external adjudicators, such as judges and arbitrators, who are concerned with the soundness of analytic procedures and fairness of the process;
- test program staff and volunteers, who want to ensure their work is protected;
- educators, who can assist by providing tips on cheating behaviors and help establish a culture of integrity among examinees; and
- examinees, who want score information as quickly as possible and want to understand the process by which scores are invalidated.

All of these stakeholder groups have different interests and understandings of how cheating affects them and the testing program. Each group might be more receptive to information presented in a different way. Testing programs might consider multiple approaches outlined in the second challenge noted previously to address each stakeholder group. Step-by-step walkthroughs of how statistics are calculated can be helpful, as can the graphical methods described in the Foley chapter.

Challenge 5: Incorporating Justice Into Our Decision-Making Frameworks

Throughout these chapters we have seen a common concern for the welfare of examinees who are false positives, and some serious concerns about the impact of false negatives. At least implicitly this discussion centers on the concept of "distributive" justice: the notion that the outcomes of a process are appropriate for the behavior of the examinee (Cropanzano, Bowen, & Gilliland, 2007). In other words, it is important that we are sanctioning the cheaters and not the honest test takers. As we consider the consequences of implementing imperfect, but useful, statistical analyses to detect cheating, we should also go a step further to consider other ways stakeholders conceptualize "justice." Given that both the psychometricians employing these methods and, even more so, outside stakeholders without this kind of expertise are generally blind to the true behavior of the examinees we flag, alternate ways we think about justice become vitally important.

"Procedural justice" is the notion that the process for determining outcomes was appropriate for the circumstances: the process was fair, unbiased, based on accurate information, that parties have had a chance to provide input into the decision, and that there is a mechanism for correcting mistakes (Cropanzano et al., 2007). These are sound principles that are repeated throughout the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, 2014), and as such should not be new ideas to test sponsors. However, it is likely that candidates rarely appreciate the full gravity of these issues prior to testing (e.g., the test sponsor has a right to cancel your score based on aberrant statistical information, the

examinee has a right to appeal those decisions), and merely including this information in a registration agreement may not be sufficient to communicate the expectations and processes. *Interactional justice* is the notion that parties are treated with dignity and respect, and that the communication process was forthright (Cropanzano et al., 2007). Testing programs often struggle to communicate effectively with their examinees, given an already tense, evaluative environment. Add to that any indication that the test sponsor suspects malfeasance on the part of the examinee, and it is even more difficult for examinees to feel their treatment has been just.

These conceptualizations of justice are correlated, but psychologically distinct, and can be empirically differentiated from the favorability of outcomes (Cropanzano et al., 2007). Expectations for what is considered “just” can vary across examinees, and from the perspective of external stakeholders. Even examinees who have cheated may feel as if they have been treated unjustly if they haven’t had an opportunity to be heard. More important, the defensibility of score invalidation decisions, both legally and in the court of public opinion, may be substantially influenced by whether candidates have been given an opportunity to present their side of the case and whether statements from the test sponsor indicate pre-judgment or are conclusory.

Our tendency as a profession is to protect ourselves from making mistakes in judgment by eliminating all judgment from the process of invalidating test scores. However, overly rigid organizational policies can backfire, making objective processes unable to consider important contextual information and therefore being unjust (Cropanzano, 2001). A better approach may be to work with other test sponsors, educational institutions, professional associations, and government agencies to promote a culture of testing integrity. To the extent that we can emphasize the extent to which behaving with integrity in testing reflects poorly on an examinee’s school, community, or profession, examinees will be more motivated to comply with our policies (Schminke, Arnaud, & Taylor, 2015).

With these ideas in mind, we can think differently about selecting cutoff values for flagging candidates. What is the process for determining whether to validate or invalidate a flagged exam record? What are the consequences? Are the outcomes appropriate given the suspected pervasiveness of cheating in the program, and the resultant likelihood of a false positive? What are the risks of validating a score from a candidate who did cheat? How clear are the expectations relating to security policies to the examinee? What incentive does the candidate have to comply with your security requests? How much information are you willing to communicate about the reasons for why a particular score was flagged and the information used to validate or invalidate a score? What is the appropriate trade-off between providing adequate due process for each candidate flagged and being able to flag candidates whose exam records show possible evidence of cheating? These are questions that come up implicitly when test sponsors examine their security policies, but rarely do we think about these issues comprehensively when we consider the overall impact of selecting analysis methods and cutoff values for detecting cheating.

CONCLUSIONS

In testing practice, data analyses provide a starting point for conducting security investigations. As such, we need to consider the methods that are best suited to detecting cheating based on the vulnerabilities of our specific testing programs, the likelihood of a cheating incident, and the ability of decision makers to digest the results. Given that

fact, we must select a range of methods that cover the most likely sources of vulnerability within our testing programs. Furthermore, we must develop a thorough understanding of how these methods are related to better utilize them for decision-making purposes.

In the process of selecting these methods, these chapters commonly raise several recommendations. First, using simulation data is extremely helpful in understanding what a security issue would look like when it occurs. By modeling multiple types, levels, and prevalence of cheating, test sponsors can select better cutoff values for security analyses.

Second, understanding how the assumptions of your statistical model interact with your testing program in terms of administration, examinees, and psychometric design are crucial to both selecting appropriate methods and interpreting results as they accumulate. In some cases, violating an assumption might be relatively inconsequential; in other cases, it might have a severe impact on the interpretability of results. Likewise, a mismatch between test design and method might also lead to problems interpreting results.

Third, there is always a trade-off between Type I and Type II errors. In few cases is this trade-off more gut-wrenching than in test security analyses. On one hand, you have the possibility of letting a cheater go unpunished, possibly issuing that person an unearned credential, and allowing the test sponsor's hard work to be undermined. On the other hand, you have possibly accusing an innocent candidate of wrongdoing, and incurring public or professional embarrassment for reaching an unsubstantiated conclusion. Again, thinking about this trade-off in the context of acting as a "just" testing organization can help test sponsors to better think through how they communicate to examinees in advance, collect evidence, and explain their decisions.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Brass, D. J., Butterfield, K. D., & Skaggs, B. C. (1998). Relationships and unethical behavior: A social networks perspective. *Academy of Management Review*, 23(1), 14–31.
- Cropanzano, R. (2001). When it's time to stop writing policies: An inquiry into procedural justice. *Human Resource Management Review*, 11(1), 31–55.
- Cropanzano, R., Bowen, D. E., & Gilliland, S. W. (2007, November). The management of organizational justice. *Academy of Management Perspectives*, 21(4), 34–48.
- Ferrara, S. (2014 October). *A framework for policies and practices to improve test security programs: PDIR*. Presentation in the annual Conference on Test Security, Iowa City, IA.
- Ferrara, S. (2016). *A framework for policies and practices to improve test security systems: Prevention, detection, investigation, and resolution (PDIR)*. Manuscript in preparation.
- Mueller, L. M., & Munson, L. (2015). Setting cut scores. In Hanvey, C. & Sady, K. (Eds.), *Practitioner's guide to legal issues in organizations* (pp. 127–162). Cham, Switzerland: Springer International.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.
- Schminke, M., Arnaud, A., & Taylor, R. (2015). Ethics, values, and organizational justice: Individuals, organizations, and beyond. *Journal of Business Ethics*, 130, 727–736.

21

THE FUTURE OF QUANTITATIVE METHODS FOR DETECTING CHEATING

Conclusions, Cautions, and Recommendations

James A. Wollack and Gregory J. Cizek

The “kitchen-sink” approach to detecting cheating (Wollack & Fremer, 2013) is—we hope—a dying paradigm. In the kitchen-sink approach, any and all analyses to identify unusual behavior are conducted, without solid understandings of the nature of the cheating suspected, the properties of the detection methods, the degrees to which they provide correlated findings, or, in many cases, accurate information about expected outcomes in null data. The approach presents a great risk for falsely labeling typical behavior as cheating or falsely attributing innocuous atypical behaviors as some form of test fraud. Wollack and Fremer (2013) suggest that the solution to the kitchen-sink approach is “to study and further develop [cheating] methodologies . . . through a combination of simulation research and real-data applications aimed at understanding the methods’ fundamental properties for a variety of situations resembling those found in practice” (pp. 8–9).

This volume represents a significant step toward that goal by bringing into better focus the breadth and quality of statistical tools available for detecting test fraud, including many newly presented methods, along with the resources needed to appropriately use and contextualize the outcomes of test fraud analyses. To help the reader quickly get up to speed on the state of research and practice, this volume includes three chapters (Chapters 2, 5, and 10) specifically aimed at the current state of affairs in cheating detection for specific types of suspected cheating behaviors. These chapters review and analyze the current approaches to detect examinees who are stealing answers from neighboring examinees, are working together during an exam, obtain live test content prior to the exam, or have their answers altered after the exam, as well as those designed to detect individual items that are compromised. Although many of the approaches described in these chapters are being used operationally, the statistical properties of some of these approaches had not been studied sufficiently well—or at all. Hence, these chapters also conduct both simulation and empirical research to help practitioners better understand how and when to use different methods.

In addition, this Handbook presents and studies 10 new methodologies for detecting cheating on tests. These new methodologies are designed to detect the wide array of cheating behaviors that most seriously jeopardize the integrity of a testing program, such as organized, group-based cheating, either through collusion, tampering, or mass distribution of secure test content. Again, the efficacy of each approach is demonstrated through both simulation and application to a real dataset, so practitioners will have the tools they need to select the best approaches for their particular test security concerns. Furthermore, by facilitating contributors' use of two common datasets throughout the volume, not only have we enhanced our understanding of the performance of existing methods and the extent to which various methods overlap, but the real datasets can serve as invaluable resources for studying future cheating detection methods and comparing results from those studies to those presented here.

Although the primary focus of this Handbook is on detecting test fraud, it is inaccurate to conceive of statistical detection methods as occurring in a vacuum. Data forensic analyses represent a piece—albeit an essential piece—of a complex puzzle, in which the complete picture might never be known. Even in those situations where parts of the picture are known, what constitutes an appropriate response to that picture is very much a matter of judgment, and there is often considerable disagreement among different stakeholders (e.g., testing programs and examinees) about the relevance or interpretation of evidence and about the appropriate conclusions. Testing programs should approach every incident of potential misconduct as though their eventual conclusions and decisions will be challenged. Consequently, every instance in which an examinee/educator/or other individual (e.g., an individual who takes a test in the place of another examinee, an individual who steals and distributes test materials) is sanctioned, including score cancelation, should be preceded by an objective and comprehensive investigation, in which both evidence in support of and evidence against misconduct are considered. For a comprehensive treatment of the entire investigative process, the reader is referred to Harris and Schoenig (2013).

In conducting an investigation, there is no substitute for using specific methods with a solid research base supporting the quality of those methods and the accuracy of the inferences they yield. However, even the best statistical evidence may not be persuasive in a legal proceeding unless it can be presented to nontechnical audiences (e.g., judges, hearing examiners, juries) in ways that are accessible and compelling. The chapters in Section III of this Handbook addressed the practical considerations associated with using statistical procedures to detect cheating on tests. We emphasize that such information—that is, detailed discussions on using and interpreting statistical analyses for cheating detection within the context of an investigation—have been noticeably lacking in the professional literature.

The first chapter in Section III (by Martineau, Jurich, Hauger, and Huff) relates to the very first part of an investigation, namely, what security vulnerabilities exist and how to best prepare for them. Martineau et al. asked this question not only of the current state of affairs, but with an eye toward where they believe the field is heading, especially as it relates to accountability testing. Their chapter focused primarily on prevention strategies, but readers of this volume might also use this chapter to ask themselves what types of statistical markers will be produced through cheating on the next generation of assessments, and how existing psychometric theory and methods can best be applied to NextGen assessments to identify NextGen misconduct.

Harris and Huang built on the ideas expressed in Martineau et al. by emphasizing that in spite of the recent advances in the statistical detection of misconduct, as the

variety of assessment formats and types of misconduct expand, our understanding of what constitutes anomalous behavior must keep pace. Depending on the particular circumstances being investigated, established indexes to detect misconduct, let alone ones that have been shown to closely follow theoretical sampling distributions, may not exist. They discussed the merits of establishing baseline data in these situations, against which to compare results from specific cases being investigated for potential fraud. Although the framework they present requires significant amounts of data and forethought to ensure that the baseline group is defined appropriately, it is eminently flexible and provides a strong foundation for beginning to study novel cheating behaviors.

In addition, this volume contains an entire chapter (by Foley) devoted to presenting results of security analyses in ways that not only accurately and fairly represent the data but also clearly illustrate the atypicality of the results. It is frequently said that a picture is worth a thousand words; we contend that in the test security world, a picture is worth much more: It is likely that whatever 1,000 words are uttered by a psychometrician to explain his or her findings may not be as well understood by the intended audience as an accurate, easily interpreted visual. In the highly technical arena of quantitative methods for cheating detection, a picture can communicate what words quite possibly cannot. Foley also emphasized the need to develop graphics to help consumers better appreciate the meaning of very small probabilities and the importance of including uncertainty estimates into our graphics. These two points are particularly salient within the test fraud context because the probabilities associated with more extreme magnitudes of cheating are often sufficiently small that people lack the context to interpret them correctly. Presenting numbers that require users to go beyond three or four decimal places is likely to strike some consumers of the data as false precision. Showing consumers information on the precision of our results communicates the science underlying such procedures and instills confidence that the results (and conclusions) can be trusted.

Continuing with this theme of communicating probabilities and accounting for uncertainty, Skorupski and Wainer (this volume) argued that communications with lay audiences are impeded by the paradigm in which current security analyses take place. Approaching cheating detection from a Bayesian perspective, they argued that traditional inferential statistics, which provide information on the likelihood of the data pattern under an assumption of no cheating, lend themselves to misinterpretations and offer little guidance into the likelihood that the specific examinee actually cheated. They demonstrated that Bayesian statistics provides an elegant framework for accounting for uncertainty—in this case, uncertainty about the base rate of cheating—and a straightforward means of evaluating the potential impact of deviations from our expected value.

Finally, effectively using quantitative methods to detect cheating requires an understanding of how those results fit into an investigation. We first addressed this notion in the introductory chapter of this volume, presenting cheating as a validity concern and stating a case for canceling scores if sufficient statistical evidence exists to question whether the score is a valid indicator of the test taker's knowledge, skill, or ability. In Section III of this book, this notion was taken one step further to consider those scenarios where a test sponsor's standards for ethics, character, and professionalism necessitate a more serious response than mere score invalidation; that is, when the reason underlying the invalid score is related to a breach of the sponsor's professional conduct requirements. Weinstein (this volume) discusses the importance of collecting collateral information before, during, and following the exam to conclude, in a legal sense, that

cheating has occurred based on the program's predetermined (and well communicated) evidentiary standards. Unsurprisingly, he argues that statistical evidence may be sufficient to cancel a score or withhold a certification, but that multiple nonquantitative forms of information related to the security incident are usually necessary to claim test fraud and impose additional sanctions.

WHERE DO WE GO FROM HERE?

Over the past 40 years, the field of test security has made tremendous strides. Many methods currently exist for detecting a wide variety of cheating behaviors. As we have seen in this volume, these approaches are quite promising and appear capable of identifying instances of test fraud for which a high proportion of the exam responses are fraudulent. Furthermore, because of the advances in detection methods, many testing programs now perform some data forensics, in either an exploratory or confirmatory manner. According to the Association of Test Publishers' 2015 Security Survey Report (ATP, 2015), 53% of testing organizations participating in the survey indicated that data forensics are an important way in which they help secure their assessments, and exactly half the organizations that reported having a security breach within the last year indicated that data forensics was one of the ways in which the breach was detected.

However, the ATP Security Survey also sheds some light on other security issues related to the use of data forensics. For example, two-thirds of the known security breaches were categorized by participants as having a low impact; that is, limited to only a single examinee and not affecting the integrity of the larger program. Only 9% of the security breaches were categorized as high impact, posing a "significant risk to [the] entire program, customer, and brand" (p. 74). The survey further found that over half the breaches resulted in needing to retire 40 or fewer items, with 43% necessitating the retirement of at most 20 items. Taken together, these findings are generally encouraging in that the magnitude of cheating in the majority of instances appears to be relatively small; on the other hand, it is perhaps somewhat discouraging that small-scale cheating situations are precisely those in which quantitative detection methods have reduced power to detect cheating.

There are also limitations of quantitative methods for detecting cheating related to the kind of cheating most often experienced in high-stakes testing contexts. For example, the use of quantitative detection methods would only aid in the detection of 6 of the 15 most common sources of breach reported in the ATP (2015, p. 66) Security Survey:

- "use of electronic devices during testing to transmit or receive answers" (#4);
- "test taker collusion during testing with other testers" (#5);
- "test taker collusion before testing (with other testers)" (#6);
- "use of live items or secure testing materials to prepare for the test" (#9);
- "unauthorized access to online exam items and distribution of bulk assessment content" (#12); and
- "test taker collusion during testing with proctors or invigilators" (#13).

The other—and arguably some of the most common—sources of breach on the ATP list refer to things such as stealing and distributing items (perhaps through social media or word-of-mouth), use of unauthorized materials, illegal coaching, stolen test materials, proxy testing, and computer hacking. For many of these, forensics can be

used to help identify the beneficiaries of these breaches, but any hope of catching the source will need to come from other security procedures.

With these contextual factors and limitations in mind, it is worthwhile to consider the future of statistical methods for detecting cheating on tests and the directions in which we would like to see the field head from both research and operational perspectives. Therefore, we conclude this book with five specific suggestions for next steps that we feel are necessary to move data forensics through its adolescence and into adulthood.

Expand the Potential for Real-Time Data Forensics

As more testing programs migrate away from paper-and-pencil formats to computer-based test administration, the opportunity exists for certain data forensics to be performed dynamically, as examinees are taking a test, to facilitate early detection. The advantages to real-time data forensics are numerous. First, it would require that programs be very clear about what they are looking for and what constitutes a suspicious anomaly, as the specific methods and flagging thresholds would need to be established a priori. Second, real-time information about suspicious behavior might also allow for on-the-fly changes to the test. As an example, suppose that, over the course of an examination, a test taker's responses were highly suggestive of preknowledge. Rather than continuing to administer items for which the examinee likely had inappropriate prior access, he or she could be automatically channeled to a more secure item block, thereby allowing his or her trait level to be estimated accurately, and perhaps even computed solely based upon the set of nonexposed items that he or she attempted. Similarly, an examinee for whom there is strong evidence of item harvesting should not continue to be shown secure items. For examinees identified as harvesters, exams could either end immediately or secure items could be replaced with decoy/retired items so as to not alert the examinee to the fact that he or she was flagged and avoid exposure of additional secure items. Finally, real-time data forensics could alert the test administrator to potential fraud, thereby allowing the proctoring team to monitor that individual more closely during the exam, in hopes of collecting collateral information.

Integrate Concern for Test Security Into the Test Development Process

Simulation data have shown that detection methods work best when examinees cheat a lot. This variable, however, is outside the control of the test sponsor. However, we have also shown that detection methods are sensitive to other factors, such as test length, item difficulty, average item latency, and the number of pilot or exposure-controlled items. As test sponsors weigh test assembly considerations, we encourage them to make their decisions with data forensics in mind. Obviously, primary consideration during test assembly should be given to ensuring fidelity to the construct being measured and coverage of the domain intended, as well as ensuring that sufficient information is available over the range of scores in which the test operates. However, that information may be corrupt for examinees who engaged in cheating. Therefore, in the interest of enhancing the validity of test score interpretations, to whatever extent possible, test specifications should be developed to facilitate detection when cheating is present.

Foster More Standardized Mechanisms for Methodological Comparisons

There have been a few data forensic challenges presented at conferences in the past couple of years (e.g., Maynes et al., 2013; Maynes, Rudner, Matthews-Lopez, & Brunner, 2013), in which teams of psychometricians perform a number of analyses on a common dataset, for purposes of determining which items and examinees may have been compromised. This book represents the first coordinated effort to not only learn about the compromise status of a dataset but to compare the effectiveness of various methodologies. To date, the literature has been surprisingly lacking with respect to comparison studies. This is particularly troubling in the area of cheating detection because “cheating” is a rather broad and not well understood area, so the particular cheating behaviors and the ways in which they are simulated (in Monte Carlo studies) or unearthed (for applied studies) vary greatly from study to study, making comparisons across studies extremely challenging. It would be of great value if the testing community took greater steps to facilitate comparisons across indexes, especially under conditions approximating those used operationally. To accomplish this, testing programs should be encouraged to make de-identified data available to researchers, especially for datasets that are believed to contain compromise. If four or five such datasets existed, each including different types and magnitudes of cheating, it would provide a rigorous and authentic proving ground for new methods, and would help guide test sponsors toward the most promising methods.

Standardization is often difficult in simulation work, because the nature of the simulation is driven, to some extent, by what the method is intended to accomplish. However, that is not to say that simulations should be designed in ways that represent best-case scenarios for the methods, which has too frequently been the case. In order for simulations to be helpful to practitioners, it is important that they reflect conditions that are typically encountered in practice. Hence, there would certainly seem to be some points where greater standardization could be achieved. We suggest that the future simulations consider, at minimum, including the simulation conditions shown in Table 21.1 which, we believe, capture the range of realistic testing conditions.

Furthermore, the testing profession does not yet know enough about how cheating behaviors manifest themselves in practice; consequently, simulation studies vary

Table 21.1 Recommended Simulation Conditions

Simulation Factor	Levels	Comment
<i>Test Length</i>	<i>n</i> = 50, 100, 250 items	Simulations should represent the likely range of test lengths, to include short, medium, and longer assessments.
<i>Sample Size</i>	<i>n</i> = 100, 500, 1000, 5000	Simulations should address testing programs that range from serving very few candidates (as in some certification fields) to many examinees.
<i>Extent of Item Compromise</i>	5%, 10%, 20%, 40%	The power of a quantitative method for detecting cheating should be demonstrated in situations where cheating ranges from minimal to substantial.
<i>Extent of Collusion</i>	2%, 5%, 10%, 20%	The power of a method for detecting collusion among examinees should be demonstrated across modest to substantial group sizes.
<i>Type I Error Level</i>	.01, .001, .0001, .00001	The alpha levels at which detection occurs should range from fairly liberal to fairly stringent.

greatly in terms of how cheating is simulated. As an example, preknowledge is often simulated under the assumption that test takers with preknowledge will get the compromised item correct with a fixed probability (often taken to be 100% or 90%), irrespective the test-taker's trait level. Other studies simulate preknowledge as a change in item difficulty, suggesting that responses to items with preknowledge are still governed by one's trait level. There is little information available to suggest which of these approaches may more closely approximate reality.

Similarly, detection methods often make assumptions about individuals with preknowledge answering questions both correctly and quickly, but very little is known about the actual response time behaviors for examinees with preknowledge. Along the same lines, item harvesting—an area of substantial concern according to the ATP Security Survey (2015)—is also something about which very little is known, resulting in a difficult environment in which to develop credible and effective detection measures. Those involved in developing and validating methods to detect cheating on tests would benefit greatly from empirical studies that attempted to describe what actual cheating looks like in practice, as such studies would allow for more realistic simulations and new and improved detection methodologies.

Develop Sound Guidance on the Use of Multiple Detection Indexes

Because different indexes are sensitive to different types of cheating or look for cheating in different ways, it is often recommended that testing entities utilize a multiple-measures approach to cheating detection (Mueller, Zhang, & Ferrara, this volume). To date, however, the field has provided little practical guidance with respect to how best to use information from multiple indexes in combination. For example, how much conceptual overlap between indexes is desirable? Commercially available programs like INTEGRITY (Castle Rock Research Corporation, 2005) and SIFT (Thompson, 2015), both of which promote multiple measures, allow for easy calculation of several different answer copying indexes. However, most of these indexes test the same research hypothesis that the number of observed answer matches exceeds its expected value, differing only with respect to how the expected number of matches is computed and whether all answers or only incorrect answers are considered. The old adage known as Segal's law says that a person with one watch always knows what time it is, whereas a person with two watches is never quite sure. There are at least four concerns that urge caution if synthesizing the results of multiple methods is contemplated.

First, concurrently conducting multiple conceptually similar methods can produce a spurious, inflated sense of convergence. In fact, we would expect the flagging patterns for two highly correlated indexes to be very similar, and in such a case, knowledge that both indexes flagged an examinee adds little to the overall evidence pattern relative to data from only one index. In contrast, evidence of cheating from two distinctly different indexes can be quite compelling.

Second, the mixing of more powerful with less powerful methods will in some cases produce differing answers that will cast doubt on cases that would seem clear-cut had only the most powerful index been used.

Third, the inflation in Type I error rate when conducting multiple statistical tests on the same data must be accounted for. Presumably, these errors can be controlled using a multiple-comparisons procedure. However, whereas each new index has some potential to either detect a new test taker who went undetected by the first index or

corroborate the results from the first index, the penalty for conducting multiple tests is a more extreme critical value, which will almost certainly lower the detection rates for all indexes involved (Wollack, 2006). Additional research is needed here to determine the best ways to increase the overall power to detect misconduct, while still keeping the false positive rate under control.

Fourth, in situations in which multiple indexes are computed, at present, there exists no way to aggregate across all statistical information to derive a single probabilistic statement, under either traditional or Bayesian approaches, about the atypicality of a particular examinee's test behavior. Such information would surely be helpful during an investigation (or legal proceeding), to help interpret data from multiple indexes, some of which were statistically significant and some of which were not.

As a result of the four issues provided above, we urge thoughtful consideration and caution regarding the conditions under which a multiple measures approach is appropriate. Although it's true that, all things being equal, it would be more desirable to have two significant indexes (correlated or not) than just one, all things are never equal. There can be a considerable cost associated with producing multiple indexes with respect to statistical power and interpretability. Indexes that disagree raise questions about the validity of the one index that identified cheating, even if the one failing to detect is known to have low power or to be sensitive to a type of cheating that was not believed to have occurred. If the criterion we proposed earlier is accepted—that all investigations should proceed as though they will eventually be challenged—then it would be wise to choose the indexes to be used very carefully. Ideally, it seems most defensible to identify, apply, and evaluate results from the single index decided upon *a priori* to be the most relevant, powerful, and appropriate. We also believe that in cases where data forensics are being conducted in a confirmatory sense (that is, after obtaining nonquantitative information regarding potential cheating), investigations should target only the specific types of cheating believed to have occurred. We assert that once a decision has been made to apply indexes that lack statistical power or are insensitive to the type of suspected cheating, the test sponsor surrenders its right to discount the relevance of those indexes, and will too often find itself in the tenuous position of attempting to explain the incongruities.

Demand Reasonable Transparency and Peer Review

The use of quantitative approaches for detecting test cheating often occurs in contexts that are considered to be high stakes. The stakes may be high for the examinee, as in contexts where the results of a test are used as part of a constellation of hurdles that an individual must pass to practice in the profession of his or her choosing; the test is often the final hurdle in an arduous, expensive, and lengthy process of professional preparation. The stakes may also be high for the organization, as in licensure contexts where an entity is charged with ensuring public health or safety, or in K-12 educational contexts where test scores are part of a system of educator and educational system evaluation.

Some quantitative methods for detecting cheating are more prevalent in practice than others but, regardless of how frequently a method is used, it would seem desirable for as much to be known about the method as possible when used in high-stakes contexts. Results of quantitative approaches often figure prominently into administrative hearings that can result in the loss of a K-12 educator's certification; statistical evidence is often considered in legal proceedings where an examinee's ability to practice his or

her chosen profession is at risk because of allegations of cheating or where a test taker faces potentially a large monetary penalty for inappropriately copying, disseminating, or disclosing protected intellectual property or other copyright violations.

Peer review is the process by which potential scientific contributions are evaluated by competent colleagues in a specialized area to determine the credibility and potential safety, efficacy, or value of, for example, a proposed new drug, a new teaching strategy, a new psychological intervention, a new insecticide, or a new counseling therapy. Proposals judged by the review of peers to be unsafe, claims that are judged to be insufficiently substantiated, findings that are the result of inadequate experimental designs, or proposals that yield only trivial benefits—or yield a benefit that is commensurate with the cost of the innovation—are typically judged by peers as not meriting publication and broader dissemination.

All of the chapters in this volume represent the good faith of their authors to describe their methods in such a way as to enable them to be scrutinized in the best traditions of peer review. The authors are to be commended for making their methods accessible for peer and public evaluation. However, not all methods currently in use, and likely not all methods that will be applied in diverse contexts in the future have been similarly evaluated. At minimum, testing programs, courts, test takers, and all those affected by the results when quantitative methods for detecting cheating are employed should demand that any approach to statistical detection of test cheating used in support of high-stakes, consequential decision making should be transparent, scrutinized by peer review, and its results should be replicable by an impartial, independent, and qualified entity. It would seem irresponsible to endorse any approach to cheating detection that fails on these basic criteria, and we encourage all future research and development efforts to take this scientific obligation seriously.

In closing, we want to affirm our conviction that the field of cheating detection is an important one. To some extent, we perceive that the field is often criticized as overidentifying innocent persons or organizations—often persons or organizations with little status or power within a system—and subjecting them to an intense and costly investigatory scrutiny that can affect their futures in profound ways. We are sensitive to that critique; indeed, we see a main value of this book as promoting methods and standards for cheating detection that are vigilant and cautious with respect to this possibility.

On the other hand, the use of sound cheating prevention and detection methods represents a significant social good. We assert—and we believe that the vast majority of the public concurs—that it is better for school children to be exposed to rigorous instruction that thoroughly prepares them for college and career opportunities than to have the test answers altered and their reported levels of learning exaggerated. We anticipate that anyone with a serious disease of the eye would prefer to be treated by a competent physician who independently passed a test of ophthalmic knowledge and skill deemed essential for safe and effective practice, than by an unsafe candidate who passed the examination because he or she obtained the questions and answers ahead of time. We suspect that most wage earners would rather have their taxes prepared by a CPA who was deemed qualified by those in the accountancy profession than by a tax preparer whose qualification was gained by copying answers from another examinee and whose lack of competence resulted in audits, fines, and penalties for the tax payer.

In short, we assert that preventing and detecting cheating is not only defensible, it is a good idea. We applaud those who take seriously their roles of public protection, certification of competence, and assurance of qualification. We hope that the methodological and procedural contributions in this volume give them additional tools to

perform those important jobs, and we look forward to further advances in the area of quantitative methods for detecting cheating that will help to further safeguard the public and honor the well-deserved, hard-earned, and legitimate accomplishments of honest test takers.

REFERENCES

- Association of Test Publishers. (2015). *Security survey report 2015*. Washington, DC: Author.
- Castle Rock Research Corporation. (2005). INTEGRITY [Computer software]. Edmonton, Alberta, Canada: Author.
- Harris, D. J., & Schoenig, R. R. W. (2013). Conducting investigations of misconduct. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 201–220). New York, NY: Routledge.
- Maynes, D., Brunnert, K., Wilson, D. J., Bontempo, B., DeLeon, C., Thomas, S., Zhang, Y., Park, J., Matthews-Lopez, J., Jones, P., & Babcock, B. (2013, October). *Potential testfraud challenge*. Presented at the annual Conference on Test Security, Madison, WI.
- Maynes, D., Rudner, L., Matthews-Lopez, J., & Brunnert, K. (2013, February). *The game's afoot: Sleuths match wits*. Presented at the annual meeting of the Association of Test Publishers, Ft. Lauderdale, FL.
- Thompson, N. (2015, November). *SIFT: Software for investigating fraud in testing*. Presented at the annual Conference on Test Security, Lawrence, KS.
- Wollack, J. A. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education*, 19, 265–288.
- Wollack, J. A., & Fremer, J. J. (Eds.). (2013). *Handbook of test security*. New York, NY: Routledge.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

APPENDIX A

Table A.1 Nominal Response Model Item Parameters Estimated From Real Dataset

	Slope Parameters Distractor				Location Parameters Distractor			
	A	B	C	D	A	B	C	D
Item 1	-0.35	0.00	0.28	0.07	-1.35	-0.43	1.64	0.14
Item 2	0.98	-0.18	-0.88	0.08	3.32	-0.18	-2.37	-0.77
Item 3	-0.02	0.70	-0.16	-0.51	-0.55	3.08	-0.25	-2.28
Item 4	0.65	0.34	-0.74	-0.25	1.59	1.01	-2.40	-0.19
Item 5	0.53	-0.27	-0.12	-0.14	1.21	-0.50	-0.39	-0.32
Item 6	-0.51	-0.10	0.43	0.18	-1.80	-0.75	2.69	-0.14
Item 7	0.63	-0.28	-0.30	-0.06	2.38	-0.35	-0.85	-1.19
Item 8	-0.84	-0.19	1.06	-0.03	-2.64	0.22	2.62	-0.20
Item 9	1.04	-0.13	1.69	-2.60	3.67	-1.04	5.22	-7.85
Item 10	0.25	-0.45	0.09	0.11	2.04	-1.45	0.14	-0.73
Item 11	-0.22	-0.08	0.25	0.04	-0.36	1.13	2.63	-3.39
Item 12	-0.24	-0.03	0.67	-0.40	-1.81	0.62	2.09	-0.89
Item 13	0.51	0.02	-0.31	-0.22	2.64	-0.58	-1.40	-0.67
Item 14	-0.77	-0.86	1.03	0.61	-2.90	-2.14	3.60	1.44
Item 15	-1.06	0.21	0.69	0.16	-2.39	-0.57	3.08	-0.12
Item 16	-0.37	0.74	-0.24	-0.13	-0.67	2.39	-0.75	-0.97
Item 17	0.99	0.14	-1.20	0.07	2.66	0.99	-4.00	0.35
Item 18	0.66	-0.01	0.11	-0.76	2.34	0.21	-0.06	-2.49
Item 19	-0.38	0.46	0.06	-0.14	-0.68	1.74	0.96	-2.02
Item 20	-0.28	-0.95	0.50	0.73	0.33	-3.66	-0.91	4.24
Item 21	-0.05	-0.10	-0.07	0.22	0.55	-0.87	-1.22	1.53
Item 22	0.43	-0.32	-0.32	0.20	2.02	-1.57	-0.51	0.07
Item 23	0.03	-0.20	-0.17	0.34	0.33	-0.13	-0.24	0.05
Item 24	0.67	0.29	-0.36	-0.59	1.59	0.05	-0.80	-0.85

(Continued)

Table A.1 (Continued)

	Slope Parameters Distractor				Location Parameters Distractor			
	A	B	C	D	A	B	C	D
Item 25	-0.05	0.62	0.07	-0.63	-0.20	2.41	-0.26	-1.94
Item 26	0.59	-0.29	-0.06	-0.25	0.24	0.21	0.03	-0.48
Item 27	-0.62	0.60	-0.10	0.13	-0.30	2.53	0.02	-2.24
Item 28	0.21	0.42	0.02	-0.65	-0.27	1.38	-0.16	-0.95
Item 29	-0.16	0.83	0.20	-0.86	-1.47	3.18	0.81	-2.52
Item 30	-0.34	0.84	0.02	-0.52	-2.18	3.51	-0.53	-0.81
Item 31	-0.39	-0.39	0.45	0.33	-0.49	-2.05	2.11	0.42
Item 32	-0.21	0.76	0.08	-0.63	-0.20	1.76	-0.43	-1.13
Item 33	-0.37	0.35	-0.11	0.14	-1.24	1.40	-1.39	1.24
Item 34	-0.27	0.29	-0.27	0.25	-0.36	-0.44	-1.14	1.94
Item 35	0.21	0.28	-0.32	-0.16	-0.67	1.69	0.27	-1.28
Item 36	0.61	-0.27	0.00	-0.34	3.34	-0.19	-0.51	-2.64
Item 37	-0.11	0.68	-0.73	0.17	-1.34	2.92	-2.44	0.86
Item 38	0.70	-0.61	0.18	-0.27	2.21	-1.66	0.10	-0.66
Item 39	0.19	-0.29	0.01	0.09	1.43	0.49	0.38	-2.30
Item 40	-0.04	-0.61	0.49	0.16	0.35	-1.31	1.19	-0.23
Item 41	-0.12	0.12	0.30	-0.30	-0.86	-0.61	1.68	-0.21
Item 42	-0.53	0.17	0.64	-0.28	-2.81	1.42	2.10	-0.71
Item 43	-0.20	0.01	-0.20	0.38	-0.29	-0.19	-1.51	1.99
Item 44	-0.30	0.54	0.13	-0.37	-2.62	1.84	1.58	-0.81
Item 45	0.58	-0.05	-0.32	-0.21	1.95	-1.56	0.29	-0.69
Item 46	0.58	0.05	-0.65	0.02	2.21	-0.68	-1.18	-0.35
Item 47	0.05	0.21	0.64	-0.89	1.19	-0.07	1.94	-3.05
Item 48	-0.42	0.07	-0.40	0.75	-1.72	0.52	-1.53	2.73
Item 49	0.75	0.07	-0.10	-0.71	3.10	0.07	-1.81	-1.37
Item 50	1.01	-0.36	-0.39	-0.26	2.57	-0.17	-0.64	-1.76
Item 51	0.37	-0.07	0.11	-0.41	1.74	-0.47	0.62	-1.89
Item 52	-0.05	-0.21	0.27	-0.01	0.40	-1.34	0.92	0.02
Item 53	0.28	0.10	0.16	-0.54	2.11	-0.26	-0.65	-1.21
Item 54	0.35	-0.37	0.71	-0.69	-0.04	-0.19	2.46	-2.23
Item 55	-0.30	-0.05	0.51	-0.16	-0.93	0.79	1.18	-1.04
Item 56	0.91	-0.25	-0.56	-0.10	2.68	-1.02	-1.52	-0.15
Item 57	0.26	-0.74	-0.02	0.49	1.44	-3.26	-0.73	2.56
Item 58	-0.02	-0.13	-0.19	0.34	-0.70	-1.64	0.85	1.49
Item 59	0.22	-0.45	1.09	-0.85	1.23	-1.02	4.33	-4.54
Item 60	-0.56	-0.17	0.19	0.53	-1.34	-0.59	-0.23	2.15
Item 61	-0.10	-1.03	1.00	0.12	1.30	-2.08	3.25	-2.48
Item 62	0.40	-0.32	-0.05	-0.03	1.82	-1.07	-0.24	-0.51
Item 63	1.08	-0.30	-1.46	0.68	3.29	-0.16	-4.97	1.84
Item 64	-0.45	0.02	0.06	0.37	-1.09	-0.22	-0.05	1.35
Item 65	0.27	-0.14	-0.04	-0.09	1.20	-1.09	0.00	-0.11

	Slope Parameters Distractor				Location Parameters Distractor			
	A	B	C	D	A	B	C	D
Item 66	0.19	0.24	-0.05	-0.39	0.22	1.37	-0.18	-1.41
Item 67	-0.52	-0.40	0.93	-0.01	-2.88	1.12	2.32	-0.57
Item 68	0.82	0.07	-0.70	-0.19	3.29	-0.24	-3.05	0.00
Item 69	0.33	0.58	-0.64	-0.27	2.37	3.61	-2.19	-3.79
Item 70	0.39	-0.10	0.07	-0.35	1.57	0.15	-0.63	-1.09
Item 71	-0.13	-0.41	-0.02	0.56	0.30	-2.13	-0.07	1.91
Item 72	0.48	0.04	-0.73	0.22	0.19	0.28	-1.43	0.95
Item 73	0.55	-0.04	0.19	-0.70	2.66	0.11	-0.05	-2.71
Item 74	-0.54	0.57	0.06	-0.09	-2.72	2.24	0.33	0.15
Item 75	-0.10	0.21	-0.17	0.07	0.35	1.90	-1.74	-0.52

Table A.2 2-PL Model Item Parameters Estimated From Real Dataset

	a	b		a	b
Item 1	0.31	-2.91	Item 39	0.31	-0.85
Item 2	1.12	-2.62	Item 40	0.53	-0.46
Item 3	0.82	-3.27	Item 41	0.45	-2.41
Item 4	0.51	-0.48	Item 42	0.50	-1.06
Item 5	0.71	-0.72	Item 43	0.52	-2.72
Item 6	0.44	-5.10	Item 44	0.43	-0.34
Item 7	0.90	-2.24	Item 45	0.85	-1.44
Item 8	1.23	-1.49	Item 46	0.73	-2.43
Item 9	0.68	-2.23	Item 47	0.56	-0.85
Item 10	0.26	-5.36	Item 48	0.82	-2.44
Item 11	0.38	-3.40	Item 49	0.89	-3.00
Item 12	0.79	-1.49	Item 50	1.34	-1.55
Item 13	0.66	-3.54	Item 51	0.32	-2.35
Item 14	0.59	-3.51	Item 52	0.33	0.33
Item 15	0.75	-3.51	Item 53	0.33	-4.87
Item 16	1.02	-2.02	Item 54	0.80	-2.19
Item 17	0.89	-1.37	Item 55	0.59	-0.17
Item 18	0.69	-2.18	Item 56	1.08	-2.09
Item 19	0.48	-1.14	Item 57	0.28	-3.49
Item 20	0.90	-3.99	Item 58	0.49	-0.77
Item 21	0.28	-2.25	Item 59	0.94	-3.15
Item 22	0.49	-2.77	Item 60	0.60	-2.70
Item 23	0.41	2.58	Item 61	1.14	-1.64
Item 24	0.76	-1.13	Item 62	0.53	-2.37
Item 25	0.68	-2.69	Item 63	0.60	-2.11
Item 26	0.79	1.04	Item 64	0.42	-1.40
Item 27	0.92	-2.03	Item 65	0.36	-1.07
Item 28	0.48	-1.32	Item 66	0.22	-2.32
Item 29	0.72	-3.07	Item 67	1.23	-0.79
Item 30	1.10	-3.02	Item 68	0.91	-2.93
Item 31	0.41	-3.06	Item 69	0.27	-4.50
Item 32	0.96	-1.19	Item 70	0.51	-1.65
Item 33	0.29	-0.01	Item 71	0.69	-1.49
Item 34	0.32	-4.39	Item 72	0.39	3.22
Item 35	0.45	-2.05	Item 73	0.55	-3.46
Item 36	0.80	-3.66	Item 74	0.58	-2.15
Item 37	0.62	-3.09	Item 75	0.27	-4.05
Item 38	0.74	-2.13			

APPENDIX B

Sample R Code for Data Manipulation and Computing Response Similarity Indices

```
#####
#           INSTALL AND LOAD THE FOLLOWING PACKAGES      #
#####  
  
install.packages("ltm")  
install.packages("psych")  
install.packages("irtoys")  
install.packages("CopyDetect")  
  
require(ltm)  
require(psych)  
require(irtoys)  
require(CopyDetect)  
  
setwd("Set your directory path here for importing datasets")  
  
#####
#           IMPORTING and MERGING the DATASET          #
#####  
  
# Import the Form 1 dataset  
  
form1 <- read.csv("Form1.csv", header=TRUE)  
  
# Exclude the flagged cases  
  
form1 <- form1[-which(form1$Flagged==1),]
```

```

# Create a new dataset with only common items from Form 1

data1 <- form1[,c(1,3,9,14,20,27,28,30,31,32,38,40,41,45,51,53,
                 56,57,64,65,69,70,72,74,77,80,81,83,86,88,90,
                 91,93,95,96,97,98,101,102,106,107,109,110,111,
                 112,114,115,117,124,127:132,134:135,137,
                 139:142,144:150,152:156,158:160,163:166,
                 168:169,171:174,176,178,179)]


# Import the Form 2 dataset

form2 <- read.csv("Form2.csv",header=TRUE)

# Exclude the flagged cases

form2 <- form2[-which(form2$Flagged==1),]

# Create a new dataset with only common items from Form 2

data2 <- form2[,c(1,3,9,12,13,17,19,22,23,25,27,29,31,35,43,45,
                 48,50,55,56,60,61,62,64,66,69,70,72,76,80,81,
                 83,84,85,86,87,88,94,95,99,100,101,103,104,
                 105,109,110,111,114,116,119,120,121,122,123,
                 125,127,129,130,131,134,135,136,138,140,142,
                 143,144,145,146,147,149,152,154,159,161,162,
                 164,165,166,167,168,170,171,172,173,174,175,
                 179,180)]


# Make the column names equal in two datasets

colnames(data2) <- colnames(data1)

# Combine two datasets with common items (Form 1 + Form 2)

data <- rbind(data1,data2)

dim(data)
head(data)

# Score the merged dataset using the key responses

key <- c(3,1,2,1,1,3,3,1,1,3,3,1,3,3,1,3,3,2,2,1,1,3,2,4,4,1,4,1,2,
        1,2,2,2,2,3,2,1,2,4,2,1,2,1,1,4,3,3,3,4,2,1,1,3,4,1,1,1,1,
        3,1,3,3,4,1,4,4,3,4,3,1,1,4,1,2,3,4,1,2,1,4,4,1,3,1,2,2,2)

dich <- data
for(i in 1:87) { dich[,i+3] = ifelse(data[,i+3]==key[i],1,0) }
dim(dich)

# Descriptive statistics for Item difficulties

describe(colMeans(dich[,4:90],na.rm=TRUE))

```

```

# Descriptive statistics for point-biserial correlations

describe(cor(cbind(rowSums(dich[,4:90],na.rm=TRUE),dich[,4:90]),
             use="pairwise.complete.obs") [2:88,1])

# Find the items with point biserial correlations < .15

rem <- which(cor(cbind(rowSums(dich[,4:90],na.rm=TRUE),
                      dich[,4:90]),
                  use="pairwise.complete.obs") [,1]<.15)

rem

# Exclude the items with low point-biserial correlations
# from nominal and dichotomous dataset

data_filtered <- data[,c(-9,-11,-21,-25,-40,-48,-58,-66,-79,-83,
                         -86,-90)]
dich_filtered <- dich[,c(-9,-11,-21,-25,-40,-48,-58,-66,-79,-83,
                         -86,-90)]

data_filtered$EID <- as.character(data_filtered$EID)
dich_filtered$EID <- as.character(dich_filtered$EID)

# Descriptive statistics for the item difficulties for final dataset

describe(colMeans(dich_filtered[,4:78],na.rm=TRUE))

# Descriptive statistics for point-biserial correlations for final
# dataset

describe(cor(cbind(rowSums(dich_filtered[,4:78],na.rm=TRUE),
                     dich_filtered[,4:78]),
             use="pairwise.complete.obs") [2:76,1])

# Key responses for the final dataset

key_filtered <- key[-rem+1]

# Estimate 2PL item parameters for the dichotomous dataset.
# These will be used later
# Note: ltm package provides almost equivalent estimates to
# IRTPRO for 1PL and 2PL models, but not for 3PL.
# If you like to use 3PL item parameters for your computation,
# you have to use IRTPRO or another software and import the
# 3PL item parameters here.

ipar.2PL <- est(resp=dich_filtered[,4:78],model="2PL",engine="ltm")
ipar.2PL$est

# Estimate 2PL ability parameters

```

```

theta.2PL <- mlebme(resp=dich_filtered[,4:78],
                      ip=ipar.2PL$est,
                      mu=0.05,sigma=1.14,method="BM") [,1]

# Below is a matrix for nominal response item parameters
# These should be estimated using another software (e.g., MULTILOG),
# and imported here

ipar.nrm <- matrix(c(
-0.39,  0.03,  0.30,  0.05, -1.36, -0.43,  1.65,  0.14,
  1.01, -0.15, -0.86,  0.00,  3.32, -0.16, -2.39, -0.77,
-0.04,  0.69, -0.13, -0.53, -0.56,  3.07, -0.24, -2.27,
  0.64,  0.32, -0.70, -0.26,  1.57,  0.99, -2.35, -0.21,
  0.52, -0.25, -0.13, -0.14,  1.22, -0.51, -0.38, -0.33,
-0.49, -0.11,  0.42,  0.18, -1.79, -0.75,  2.70, -0.15,
  0.66, -0.27, -0.31, -0.08,  2.38, -0.33, -0.84, -1.21,
-0.81, -0.16,  1.08, -0.11, -2.60,  0.22,  2.61, -0.22,
  0.54, -0.51,  1.19, -1.22,  2.53, -2.10,  4.06, -4.49,
  0.29, -0.54,  0.11,  0.14,  2.05, -1.45,  0.14, -0.73,
-0.25, -0.08,  0.26,  0.06, -0.38,  1.14,  2.65, -3.41,
-0.29, -0.01,  0.70, -0.40, -1.83,  0.62,  2.09, -0.87,
  0.52,  0.07, -0.32, -0.27,  2.64, -0.56, -1.41, -0.68,
-0.82, -0.76,  1.01,  0.57, -2.91, -1.91,  3.49,  1.33,
-1.03,  0.18,  0.71,  0.14, -2.28, -0.59,  3.04, -0.17,
-0.38,  0.76, -0.25, -0.12, -0.71,  2.37, -0.76, -0.90,
  1.01,  0.16, -1.28,  0.11,  2.66,  1.00, -4.03,  0.37,
  0.70,  0.00,  0.10, -0.80,  2.35,  0.20, -0.06, -2.48,
-0.38,  0.47,  0.07, -0.16, -0.67,  1.72,  0.95, -1.99,
-0.26, -0.81,  0.29,  0.78,  0.32, -3.62, -0.91,  4.20,
-0.02, -0.15, -0.07,  0.24,  0.56, -0.85, -1.22,  1.52,
  0.45, -0.37, -0.31,  0.23,  2.02, -1.59, -0.49,  0.07,
  0.04, -0.19, -0.18,  0.32,  0.32, -0.12, -0.24,  0.04,
  0.66,  0.24, -0.34, -0.56,  1.57,  0.04, -0.80, -0.81,
-0.04,  0.62,  0.08, -0.65, -0.22,  2.40, -0.26, -1.92,
  0.61, -0.30, -0.06, -0.25,  0.24,  0.22,  0.03, -0.49,
-0.60,  0.58, -0.11,  0.13, -0.26,  2.52,  0.03, -2.29,
  0.18,  0.43,  0.03, -0.64, -0.25,  1.37, -0.16, -0.95,
-0.14,  0.84,  0.18, -0.89, -1.46,  3.18,  0.80, -2.52,
-0.38,  0.83,  0.03, -0.48, -2.08,  3.43, -0.56, -0.79,
-0.36, -0.42,  0.46,  0.33, -0.47, -2.04,  2.11,  0.40,
-0.23,  0.77,  0.09, -0.64, -0.20,  1.76, -0.43, -1.12,
-0.39,  0.37, -0.15,  0.17, -1.22,  1.40, -1.41,  1.22,
-0.30,  0.30, -0.28,  0.27, -0.34, -0.44, -1.16,  1.93,
  0.21,  0.29, -0.33, -0.17, -0.68,  1.69,  0.28, -1.29,
  0.63, -0.31, -0.01, -0.31,  3.35, -0.17, -0.52, -2.66,
-0.11,  0.69, -0.72,  0.15, -1.35,  2.91, -2.38,  0.83,
  0.69, -0.57,  0.15, -0.27,  2.18, -1.60,  0.07, -0.65,
  0.19, -0.29,  0.03,  0.07,  1.45,  0.49,  0.37, -2.31,
-0.04, -0.60,  0.46,  0.18,  0.34, -1.28,  1.18, -0.24,
-0.17,  0.12,  0.32, -0.28, -0.87, -0.61,  1.67, -0.19,

```

```

-0.51,  0.16,  0.64, -0.29, -2.80,  1.41,  2.08, -0.68,
-0.22,  0.00, -0.18,  0.40, -0.31, -0.19, -1.49,  1.99,
-0.28,  0.53,  0.14, -0.40, -2.64,  1.84,  1.59, -0.79,
  0.60, -0.05, -0.33, -0.22,  1.95, -1.56,  0.30, -0.70,
  0.59,  0.07, -0.62, -0.04,  2.19, -0.68, -1.12, -0.38,
  0.03,  0.17,  0.63, -0.83,  1.17, -0.11,  1.91, -2.96,
-0.55,  0.10, -0.33,  0.78, -1.76,  0.52, -1.49,  2.74,
  0.80,  0.11, -0.20, -0.71,  3.12,  0.08, -1.87, -1.33,
  1.05, -0.39, -0.34, -0.33,  2.56, -0.19, -0.57, -1.81,
  0.40, -0.07,  0.16, -0.48,  1.73, -0.45,  0.63, -1.90,
-0.06, -0.21,  0.27,  0.00,  0.41, -1.33,  0.91,  0.01,
  0.29,  0.09,  0.15, -0.54,  2.11, -0.27, -0.64, -1.20,
  0.35, -0.31,  0.74, -0.77, -0.03, -0.13,  2.46, -2.31,
-0.29, -0.04,  0.50, -0.16, -0.92,  0.78,  1.19, -1.05,
  0.89, -0.26, -0.52, -0.10,  2.63, -1.01, -1.46, -0.16,
  0.30, -0.80, -0.04,  0.54,  1.41, -3.23, -0.72,  2.54,
-0.04, -0.11, -0.18,  0.34, -0.71, -1.62,  0.83,  1.50,
  0.35, -0.32,  1.23, -1.25,  1.38, -0.83,  4.46, -5.01,
-0.58, -0.15,  0.22,  0.52, -1.35, -0.54, -0.24,  2.13,
  0.03, -0.95,  1.10, -0.18,  1.37, -2.05,  3.28, -2.60,
  0.42, -0.31, -0.06, -0.05,  1.81, -1.06, -0.24, -0.51,
  1.06, -0.34, -1.38,  0.66,  3.19, -0.14, -4.79,  1.74,
-0.43,  0.01,  0.05,  0.38, -1.06, -0.24, -0.04,  1.34,
  0.28, -0.17, -0.04, -0.08,  1.20, -1.10,  0.00, -0.11,
  0.19,  0.25, -0.04, -0.41,  0.23,  1.38, -0.19, -1.42,
-0.53, -0.37,  0.94, -0.04, -2.89,  1.15,  2.31, -0.57,
  0.88,  0.11, -0.86, -0.12,  3.30, -0.27, -3.07,  0.04,
  0.33,  0.58, -0.70, -0.21,  2.37,  3.63, -2.23, -3.78,
  0.40, -0.08,  0.05, -0.37,  1.57,  0.15, -0.63, -1.08,
-0.13, -0.44, -0.01,  0.58,  0.30, -2.15, -0.07,  1.92,
  0.48,  0.03, -0.72,  0.21,  0.19,  0.27, -1.39,  0.93,
  0.54, -0.06,  0.20, -0.67,  2.65,  0.10, -0.06, -2.69,
-0.50,  0.55,  0.06, -0.10, -2.66,  2.22,  0.29,  0.15,
-0.12,  0.21, -0.17,  0.08,  0.36,  1.90, -1.74, -0.52),
nrow=75,ncol=8,byrow=TRUE)

#####
# CREATE the NULL DISTRIBUTIONS for ANGOFF's B and SAUPE's INDICES #
#####
# 2,000 pairs of response vectors will be generated by randomly
# sampling examinees from different test centers and matching
# Then the following quantities will be computed for each pair
# of response vectors
#
#   W_ij: the number of items examinee i answer incorrectly
#   w_j: the number of items examinee j answer incorrectly
#   W_ij: the number of items both examinees answer incorrectly
#   w_ij: the number of items both examinees have identical
#         incorrect responses
#
#   R_i: the number of items examinee i answer correctly
#   R_j: the number of items examinee j answer correctly
#   R_ij: the number of items both examinees answer correctly

```

```

# If you like, you can increase the number of pairs as high as
# you want. Initial exploration suggested that 2,000 pairs is
# more than enough to get precise estimates for the regression
# weights needed to compute Saupe's index and Angoff's B

null.dist <- as.data.frame(matrix(nrow=2000, ncol=9))
colnames(null.dist) <- c("ID_i", "ID_j", "W_i", "W_j", "W_ij", "w_ij",
                         "R_i", "R_j", "R_ij")

# the following loop can take a long time

for(i in 1:2000) {

  # Randomly select an examinee EID

  a = sample(data_filtered$EID,1)
  center_i    = data_filtered[which(data_filtered==a),]$cent_id
  b = sample(
    data_filtered[which(data_filtered$cent_id!=center_i),]$EID,1
  )

  null.dist[i,1:2] <- c(a,b)

  resp.a <- data_filtered[which(data_filtered==a),4:78]
  resp.b <- data_filtered[which(data_filtered==b),4:78]

  scored.a <- dich_filtered[which(dich_filtered==a),4:78]
  scored.b <- dich_filtered[which(dich_filtered==b),4:78]

  null.dist[i,3] <- sum(scored.a==0)
  null.dist[i,4] <- sum(scored.b==0)
  incorrect.items <- which(scored.a==0 & scored.b==0)
  null.dist[i,5] <- length(incorrect.items)
  if(null.dist[i,5]!=0){
    null.dist[i,6] <- sum(resp.a[incorrect.items]==
                           resp.b[incorrect.items])
  } else null.dist[i,6]=0

  null.dist[i,7] <- sum(scored.a==1)
  null.dist[i,8] <- sum(scored.b==1)
  null.dist[i,9] <- length(which(scored.a==1 & scored.b==1))

}

null.dist$Ri_Rj <- null.dist$R_i*null.dist$R_j
null.dist$Wi_Wj <- null.dist$W_i*null.dist$W_j

head(null.dist)

# Saupe's method

# Scatterplot: w_ij vs W_ij

```

```

plot(null.dist$W_ij,null.dist$w_ij)
cor(null.dist$W_ij,null.dist$w_ij,use="pairwise.complete.obs")

# Scatterplot: R_ij vs R_i*R_j

plot(null.dist$R_i*null.dist$R_j,null.dist$R_ij)
cor(null.dist$R_i*null.dist$R_j,null.dist$R_ij,
    use="pairwise.complete.obs")

# Regression weights for incorrect responses

beta0_w <- coef(lm(null.dist$w_ij~null.dist$W_ij))[1]
beta1_w <- coef(lm(null.dist$w_ij~null.dist$W_ij))[2]
betas_w <- summary(lm(null.dist$w_ij~null.dist$W_ij))$sigma

# Regression weights for correct responses

beta0_r <- coef(lm(null.dist$R_ij~null.dist$Ri_Rj))[1]
beta1_r <- coef(lm(null.dist$R_ij~null.dist$Ri_Rj))[2]
betas_r <- summary(lm(null.dist$R_ij~null.dist$Ri_Rj))$sigma

# Angoff's B

# Original Angoff's B uses stratification
# But, here a similar regression approach will be used as
# in the Saupe's method. Only difference is the predictor
# used in the regression equation

# Scatterplot: w_ij vs W_ij

plot(null.dist$Wi_Wj,null.dist$w_ij)
cor(null.dist$Wi_Wj,null.dist$w_ij,use="pairwise.complete.obs")

# Regression weights for Angoff's B

angoff_b0 <- coef(lm(null.dist$w_ij~null.dist$Wi_Wj))[1]
angoff_b1 <- coef(lm(null.dist$w_ij~null.dist$Wi_Wj))[2]
angoff_b <- summary(lm(null.dist$w_ij~null.dist$Wi_Wj))$sigma

# These regression weights are later used for answer copying
# detection

#####
# COMPUTING PERSON-FIT INDICES and RESPONSE SIMILARITY #
# INDICES FOR A PAIR OF EXAMINEES                      #
#####

detach("package:psych", unload=TRUE)

# Suppose you want to compute the person-fit indices and
# response similarity indices for the examinees with ID numbers
# e100019 and e101327 in test center 5302
# Row number for examinee e100019

```

```

i=which(dich_filtered$EID=="e100019")

# Row number for examinee e101327

j=which(dich_filtered$EID=="e101327")

# Using dichotomous response outcomes

# i is treated as copier and j is treated as source

fit <- CopyDetect1(data = dich_filtered[,4:78],
                     item.par=ipar.2PL$est,pair = c(i,j))

# The followings are p-values from response similarity indices

fit$W.index$p.value
fit$GBT.index$p.value
fit$K.index$k.index
fit$K.variants$K1.index
fit$K.variants$K2.index
fit$K.variants$S1.index
fit$K.variants$S2.index

# Compute Person Fit Indices

# H statistic

sig.ij <- cov(t(dich_filtered[,4:78]),
                 use="pairwise.complete.obs")
diag(sig.ij) <- rep(0,nrow(dich_filtered))
H.num <- apply(sig.ij,1,sum)
p.cor.n <- rowSums(dich_filtered[,4:78],na.rm=TRUE)/75
p.cor.i <-
  colSums(dich_filtered[,4:78],na.rm=TRUE)/nrow(dich_filtered)
p.cor.i.matrix <- t(matrix(p.cor.i,75,nrow(dich_filtered)))
H.denom1 <- matrix(p.cor.n,nrow(dich_filtered),1) %*%
  matrix(1-p.cor.n,1,nrow(dich_filtered))
H.denom2 <- matrix(1-p.cor.n,nrow(dich_filtered),1) %*%
  matrix(p.cor.n,1,nrow(dich_filtered))
H.denom <- ifelse(H.denom1>H.denom2,H.denom2,H.denom1)
diag(H.denom) <- rep(0,nrow(dich_filtered))
H.denom <- apply(H.denom,1,sum)
H <- H.num/H.denom

H[i]

# D statistic

obs <- dich_filtered[i,4:78]
pred <- irf(ip=ipar.2PL$est,x=theta.2PL[i])$f
mean((obs-pred)^2)

```

```

# Modified lz statistic (de la torre & Deng, 2008)
# Step numbers are aligned with the steps described in
# the paper
vector.c <- dich_filtered[i,4:78]

# Step 1

th.eap <- eap(dich_filtered[i,4:78],ip=ipar.2PL$est,
                 qu=normal.qu(n=40))

# Step 2

th <- th.eap[,1]*(1+th.eap[,2]^2)

# Step 3

prob.c <- irf(ip=ipar.2PL$est,x=th)$f
prob.q <- 1-prob.c
10 <- sum((vector.c*log(prob.c))+((1-vector.c)*log(prob.q)))
exp.lz <- sum((prob.c*log(prob.c))+(prob.q*log(prob.q)))
var.lz <- sum(prob.c*(1-prob.c)*(log(prob.c/(1-prob.c))^2))
lz <- (10-exp.lz)/sqrt(var.lz)

# Step 4

th.s <- rnorm(1000,mean=th,sd=1/(1+tf(ipar.2PL$est,x=th)$f))

# Step 5

resp.tmp <- sim(ip=ipar.2PL$est,x=th.s)

# Step 6i: Repeat Step 1 and Step 2 for each new theta

th.s.eap <- eap(resp.tmp,ip=ipar.2PL$est,qu=normal.qu(n=40))
th.s.up <- th.s.eap[,1]*(1+th.s.eap[,2]^2)

# Step 7

prob.c.s <- irf(ip=ipar.2PL$est,x=th.s.up)$f
prob.q.s <- 1-prob.c.s
10.s <- rowSums((resp.tmp*log(prob.c.s))+((1-resp.tmp)*
                  log(prob.q.s)))
exp.lz.s <- rowSums((prob.c.s*log(prob.c.s))+(prob.q.s*
                  log(prob.q.s)))
var.lz.s <- rowSums(prob.c.s*(1-prob.c.s)*
                  (log(prob.c.s/(1-prob.c.s))^2))
lz.s <- (10.s-exp.lz.s)/sqrt(var.lz.s)

# Associated p value

length(which(lz.s<=lz))/1000

```

```

# Using nominal response outcomes

data.tmp <- data_filtered[,4:78]

for(kkk in 1:75) {data.tmp[,kkk] <- as.character(data.tmp[,kkk])}

fit2 <- CopyDetect2(data = data.tmp,
                      item.par=ipar.nrm,
                      pair = c(i,j),
                      options=c("1","2","3","4"))

fit2$W.index$p.value
fit2$GBT.index$p.value
fit2$K.index$k.index
fit2$K.variants$K1.index
fit2$K.variants$K2.index
fit2$K.variants$S1.index
fit2$K.variants$S2.index

# Angoff's B and Saupe's t

www <- length(which(dich_filtered[i,4:78]==0 &
                      dich_filtered[j,4:78]==0))

if(www!=0) {

  Ri_Rj = sum(dich_filtered[i,4:78]==1)*
          sum(dich_filtered[j,4:78]==1)

  Wi_Wj = sum(dich_filtered[i,4:78]==0)*
          sum(dich_filtered[j,4:78]==0)

  common.incorrect <- which(dich_filtered[i,4:78]==0 &
                               dich_filtered[j,4:78]==0)

  w_ij = sum(data_filtered[i,common.incorrect]==
              data_filtered[j,common.incorrect])

  common.correct <- which(dich_filtered[i,4:78]==1 &
                           dich_filtered[j,4:78]==1)

  # Returns the p value for Angoff's B

  t_b <- (w_ij - (angoff_b0+angoff_b1*Wi_Wj))/angoff_b
  1-pt(t_b,1998)

  # Returns the p value for Saupe's t

  t_w <- (w_ij - (beta0_w+beta1_w*length(common.incorrect)))/
          betaS_w
  t_r <- (length(common.correct) - (beta0_r+beta1_r*Ri_Rj))/
          betaS_w
  (1-pt(t_w,1998))*(1-pt(t_r,1998))

}

```

APPENDIX C

Openbugs Code for Fitting the Bayesian HLM and Estimating Growth Aberrance

```
model {

#Specifies Multivariate Normal Likelihood of data (y)
#y is a Nx2 data matrix of scale scores (N examinees by 2 time
points)
#Means (mu) are nx2 (n groups by 2 time points)
#Precision matrix (Tau) of scores over Time is Group specific
#Grouping variable (g=1,2,. . . ,n) indicates which individual
is in which group

for (i in 1:N) {
y[i,1:2] ~ dmnorm(mu[g[i],1:2],Tau[g[i],1:2,1:2])
}

#Specify noninformative prior (Rho is 2x2 identity matrix) for
Tau
#Tau is Wishart distributed, with df=group size-1: G_N[j]-1
#This section also converts Tau into V/C (SIG) and Correlation
(R) matrices
for (j in 1:n) {
df[j] <- G_N[j]-1
Tau[j,1:2,1:2] ~ dwish(Rho[,],df[j])
INV[j,1:2,1:2] <- inverse(Tau[j,1:2,1:2])
    for (k in 1:2) {
        for (l in 1:2) {
            SIG[j,k,l] <- 2*INV[j,k,l]
            R[j,k,l] <- SIG[j,k,l]/ sqrt(SIG[j,k,k]*SIG[j,l,l])
        }
    }
}
```

```
#group-by-time means are nested within time points
for (j in 1:n) {
    for (t in 1:2) {
        mu[j,t] ~ dnorm(MuT[t],TauT[t])
    }
}

#Specify noninformative priors for Time effects
for (t in 1:2) {
    MuT[t] ~ dnorm(0,0.0001)
    TauT[t] ~ dgamma(0.1,0.1)
    VarT[t] <- 1/TauT[t]
}

#Logical node calculates Growth Aberrance (GA) conditional on
#Group variability
for (j in 1:n) {
    GA[j] <- (mu[j,2]-MuT[2])/sqrt(SIG[j,2,2]) - (mu[j,1]-MuT[1])/sqrt(SIG[j,1,1])
}
```

CONTRIBUTORS

Dmitry Belov is Senior Research Scientist at the Department of Psychometric Research of the Law School Admission Council. His research interests include automated test assembly and item pool analysis, detection of group cheating, and applications of natural language processing in test development and psychometrics. Dr Belov earned a PhD in Computer Science from the Institute of Engineering Cybernetics of the National Academy of Sciences of Belarus.

Scott Bishop holds a PhD in Educational Measurement and Statistics. He currently serves as Lead Psychometrician in the Center for Educational Testing and Evaluation (CETE) at the University of Kansas. During his career, he has provided technical support to large-scale norm- and criterion-referenced testing programs, consulted on formative assessment programs, supported the development of several off-the-shelf tests, and provided oversight for a number of special linking and validity studies. Dr Bishop has authored research publications and presentations on equating, vertical scaling, and erasure analysis. He is a former Editor of the *NCME Newsletter*.

Keith A. Boughton is a Senior Research Scientist at Data Recognition Corporation. His professional experience has focused on benchmark and summative assessment programs. His research interests include cheating detection, adaptive testing, test speededness, Markov Chain Monte Carlo algorithms (MCMC), hierarchical IRT models, and multidimensional IRT models.

Gregory J. Cizek is Guy B. Phillips Distinguished Professor of Educational Measurement and Evaluation at the University of North Carolina–Chapel Hill. His research focuses on standard setting, test security, validity, and testing policy. Previously, he managed national licensure and certification programs, worked on a state assessment program, and he began his career as an elementary school teacher. He has served as President of the National Council on Measurement in Education (NCME), member of the National Assessment Governing Board, and Secretary of Division D and the Professional Licensure and Certification Special Interest Group of the American Educational Research Association (AERA).

J. Michael Clark is a Senior Analytical Consultant in the SAS EVAAS for K-12 department at SAS Institute Inc. His research interests include growth, value-added modeling, data forensics, and innovative methods for measurement, visualization, and reporting. He has provided psychometric services in both educational and professional testing in previous roles at Pearson and Applied Measurement Professionals, Inc. He is an active member of numerous professional organizations, including AERA, NCME, and the Psychometric Society.

Phil Dickison is the Chief Officer, Examinations for National Council of the State Boards of Nursing (NCSBN). Dr Dickison oversees the development, administration and psychometrics of the NCLEX and NNAAP™/MACE™ examinations. His research interests focus on computer adaptive testing, security, and measurement of clinical decisions ability. Previously, he served as the Associate Director of the National Registry of Emergency Medical Technicians. He earned his PhD in Quantitative Research in Evaluation and Measurement in Education from The Ohio State University.

Carol A. Eckerly is a Psychometrician at Alpine Testing Solutions. Her current research focuses on examination security. She earned her PhD in Educational Psychology—Quantitative Methods from the University of Wisconsin—Madison, where she taught courses in statistical methods, experimental design, and factor analysis.

Karla Egan is Founder and Principal at EdMetric LLC, where she works with state and national practitioners and policymakers to design, evaluate, and improve assessment systems and practices. During her career, she has assisted states in the development of summative assessments; she has made presentations and published research on issues related to assessment design, test security, standard setting, and test security. Her accomplishments include creating an innovative framework for achievement level descriptors. Dr Egan received her PhD from the University of Massachusetts, Amherst.

Steve Ferrara is Senior Advisor for Measurement Solutions at Measured Progress. Dr Ferrara conducts psychometric research and designs large scale and formative assessments for K-12 educational achievement, special education, and English language proficiency assessment programs. He was co-recipient, AERA Cognition and Assessment Special Interest Group 2014 award for Outstanding Contribution to Practice in Cognition and Assessment; and co-recipient, AERA Division D 2006 award for Significant Contributions to Educational Measurement and Research Methodology. In recent publications and presentations, Dr Ferrara has focused on prevention of security threats and detection, investigation, and resolution of possible test security violations.

Joe Fitzpatrick is a graduate student in the department of Educational Psychology at the University of Kansas. His primary research interests include standard setting, equating, and test security.

Brett P. Foley is Director of Professional Credentialing and a Senior Psychometrician at Alpine Testing Solutions. He has worked with many types of testing programs in licensure, certification, and education. Dr Foley received his PhD in Quantitative,

Qualitative, and Psychometric Methods from the Department of Educational Psychology at the University of Nebraska–Lincoln. He is the currently Website Content Editor for NCME and is a past president of the Northern Rocky Mountain Educational Research Association. His research interests include standard setting, policy considerations in testing, and using visual displays to inform the test development process.

Matthew Gaertner is a Principal Research Scientist at SRI International. His methodological interests include multilevel models, categorical data analysis, and Item Response Theory. Substantively, his research focuses on the effects of educational policies on access, persistence, and achievement. Prior to joining SRI, he held positions in Pearson's Research & Innovation Network and at American Institutes for Research. Dr Gaertner received a Spencer Foundation Dissertation Fellowship and the 2013 and 2011 Best Paper Awards from the Association for Institutional Research. Dr Gaertner earned a PhD in research and evaluation methodology from the University of Colorado–Boulder.

Deborah J. Harris is Vice President of Measurement Research at ACT, Inc. Dr Harris has more than 30 years of experience in equating, scaling, and issues related to large-scale, high-stakes testing, such as context effects, domain scoring, and forensics research/cheating detection. She is also an adjunct faculty member at the University of Iowa, and is the author or co-author of more than 140 psychometric-related publications and presentations.

Jeffrey B. Hauger is the Director of Assessments at the New Jersey Department of Education. He has a doctorate of education in research and evaluation methods from the University of Massachusetts, Amherst. Over the past 5 years, Dr. Hauger has worked on the Partnership for Assessments of Readiness for College and Careers (PARCC) where he served as the co-chair of PARCC's research and psychometric working group. He is currently co-chair of the state leadership working group that consists of assessment directors from all PARCC states that make policy and testing recommendations to PARCC's Governing Board.

Chi-Yu Huang is a Principal Psychometrician in the Measurement Research department at ACT, Inc., in Iowa City, Iowa. Dr Huang specializes in research and analysis related to equating and scaling, educational psychometrics, data forensics, and test accommodations.

Kristen Huff is Vice President, Assessment and Research at Curriculum Associates. Her work focuses on ensuring the coherence of assessment design, interpretation, use, and policy to advance equity and high-quality education for all students. Previously, she worked on designing statewide tests used for accountability as well as college placement and admission tests. Dr Huff received her EdD in Measurement, Research, and Evaluation Methods from the University of Massachusetts–Amherst in 2003.

Daniel Jurich is a Psychometrician at the National Board of Medical Examiners where his responsibilities include managing the psychometric activities for various licensure and in-training examinations. His primary research interests include

improving the diagnostic utility of assessments to aid in tailoring remediation and data forensic techniques to examine test security. Previously, he was a Fellow at the Regents Research Fund where he worked on a statewide assessment program with a focus on evaluating test security methods and policies. He is an active member of NCME and AERA.

Doyoung Kim, is a Senior Psychometrician at National Council of State Boards of Nursing (NCSBN). His current research focuses on model-data fit, dimensionality, and fairness in testing. Previously, he was a Senior Psychometrician at American Institutes for Research (AIR). He earned his PhD in Quantitative, Qualitative, and Psychometric Methods (QQPM) from the University of Nebraska-Lincoln.

Joseph A. Martineau is a Senior Associate with the Center for Assessment, located in Dover, New Hampshire. His work focuses on improving assessment practice at the intersection of policy, psychometrics, and operations with an emphasis on dimensionality, growth modeling, and accountability. Previously, he served as Psychometrician, State Assessment & Accountability Director, and Deputy Superintendent in the Michigan Department of Education responsible for teacher licensure, K-12 assessment, K-12 accountability, and educator evaluation. He has also been a member of the board of directors of NCME and a co-chair of the Smarter Balanced Assessment Consortium.

Dennis D. Maynes is Chief Scientist at Caveon Test Security. He has pioneered several methods for the statistical detection of potential test fraud, including the use of clusters to detect cheat rings and the use of embedded verification tests to detect brain-dump users. He has conducted more than 450 data forensics projects for more than 50 organizations, including 11 state departments of education, 10 medical programs, and 12 information technology certification programs. Maynes holds a Master's Degree in statistics from Brigham Young University.

Yuanyuan (Malena) McBride is a Research Scientist at Pearson. She received her BS in Earth Sciences from Nanjing University in China, an MS in Geology from Baylor University, and a PhD in Research, Measurement, and Statistics from Texas A&M University. Her primary research interests include student growth measurement, statistical detection of test fraud, and matching methods for causal inference.

Lorin Mueller is Managing Director of Assessment for the Federation of State Boards of Physical Therapy, where he oversees the development of the National Physical Therapy Examinations. Prior to joining FSBPT in 2011, he was a Principal Research Scientist at the American Institutes for Research, where he worked for 11 years in high-stakes selection, assessment development, and career preparation research. Dr Mueller has a PhD in Industrial and Organizational Psychology from the University of Houston.

Stephen Murphy is the Director of Research at Houghton Mifflin Harcourt. His research specializations include psychometric and research for Educational and Clinical Assessments comprising norms development, IRT calibration, scaling, and equating, adaptive testing, standard setting, growth and development, reliability and validity studies, and guidance on educational and clinical policy. Previously, he was a director at Pearson

where he served as the leader of a team of psychometricians supporting numerous high stakes K-12 assessments. In this role, he lead numerous state testing contracts and conducted research in scaling and equating, standard settings, educational policy, and test security.

Lisa S. O'Leary is a Senior Psychometrician at Alpine Testing Solutions, Inc., where she provides consultation and analyses to certification programs on their program design, exam development and maintenance processes, and security plans. Her research interests include comprehensive approaches to test security for prevention, detection, and enforcement, item and exam stability over time, and exam development solutions for domains with rapidly changing content. She has presented at national and international conferences and to multiple organizations on research findings related to test fraud. Previously, she worked in assessment and evaluation at MIT and Tufts University.

Hao Ren is a Senior Research Scientist at Pacific Metrics Corporation. His research focuses on Bayesian methodology, computational statistics, and computerized adaptive testing (CAT). Previously, he worked as Research Scientist at Data Recognition Corporation and CTB/McGraw-Hill Education, with experience on various operational projects and research topics.

William P. Skorupski is an Associate Professor in the department of Educational Psychology at the University of Kansas and Co-Coordinator of the Research, Evaluation, Measurement, and Statistics program. His research focuses on applications of Item Response Theory and Bayesian statistics for solving practical measurement problems, including scaling, parameter estimation, standard setting, and test security.

Jessalyn Smith is a Research Scientist with Data Recognition Corporation. Her experience and expertise are in the areas of statistical consulting and formative and summative assessment practices. Her interests include statistical computing, latent class modeling, statistical methods applied to assess multidimensionality, and item and person compromise.

Russell W. Smith is a Senior Psychometrician and the Director of IT Credentialing at Alpine Testing Solutions, Inc. He is primarily responsible for conducting and overseeing the quality of test development and psychometric analyses including classical test theory, item response theory, Rasch, equating, scaling and standard setting. His research interests include psychometric security analyses and automated test assembly algorithms. He has presented extensively to professional organizations and at national conferences, and he co-authored a chapter in the *Handbook of Test Development* (2006). Prior to joining Alpine Testing Solutions, Russell was a Senior Psychometrician at Prometric and Galton Technologies.

Howard Wainer is Distinguished Research Scientist at the National Board of Medical Examiners. He is a Fellow of AERA and the American Statistical Association. Among his honors, he has received the ACT/AERA E. F. Lindquist Award for Outstanding Research in Testing & Measurement (2015), the Psychometric Society Lifetime Achievement Award (2013), the Samuel J. Messick Award for

Distinguished Scientific Contributions from Division 5 of the American Psychological Association (2009), and the NCME Career Achievement Award for Contributions to Educational Measurement (2007). His most recent book is *Truth or Truthiness: Distinguishing Fact from Fiction by Learning to Think Like a Data Scientist* (Cambridge, 2016).

Marc J. Weinstein is Chief Privacy Officer and Vice President of Investigative Services for Caveon, LLC, and an attorney with his own law practice. In his law practice, he counsels testing organizations regarding trade secret protection, copyright law, program policies, stakeholder agreements, and test security violations. He has broad experience conducting all aspects of cheating and test-theft investigations throughout North America, across every sector of testing, including significant experience in high profile investigations in certification examinations, admissions tests, and K-12 assessments. Marc began his career in 1998 as a prosecutor in the Office of the State Attorney for Miami-Dade County, Florida.

James A. Wollack is a Professor of Quantitative Methods in the Educational Psychology Department and the Director of Testing & Evaluation Services and the Center for Placement Testing at the University of Wisconsin-Madison. His research focuses on test security, item response theory, test construction, and placement testing. He is currently on the Governing Council for the National College Testing Association and the Executive Committee for the Conference on Test Security, and has served on the NCME Board of Directors and as past-chair of the Measurement Services Special Interest Group of AERA.

Ada Woo is the Director of Measurement and Testing at National Council of State Boards of Nursing (NCSBN). In this position, she is responsible for the management all NCSBN examination programs and related services. She has more than 10 years of experience in the licensure and certification testing field. Prior to joining NCSBN, Dr Woo was a part of the psychometric team at the Federation of State Boards of Physical Therapy, where she participated in the test development of the National Physical Therapy Examinations. She holds a doctorate in Quantitative Psychology from the University of Texas at Arlington.

Yu Zhang is a Senior Psychometrist at the Federation of State Boards of Physical Therapy. His work and research focuses on development and operational aspects of licensure examinations. Previously, he was a Senior Analyst at American Institutes for Research where he worked on several projects of K-12 or adult literacy assessment. He has given several presentations on the topic of test security at the annual meetings of NCME, AERA, and the Conference on Test Security.

Cengiz Zopluglu is an Assistant Professor of Research, Measurement, and Evaluation in the School of Education at the University of Miami. His research interests are focused on three primary areas: item response theory, test score integrity, and nonlinear mixed-effects models. He received a PhD from University of Minnesota in 2013 and began his career as a middle-school math teacher in Turkey.

INDEX

Page numbers in italic format indicate figures and tables.

- aberrant examinees 164–5, 167–74
aberrant growth 232, 233, 236, 237, 243
aberrant response patterns: approaches for 73–92; defined 72; identifying and investigating 72–3; introduction to 25–6; lognormal RT model and 181, 183–4; machine-learning for 92–5
accountability testing: conclusion about 391; human behavior and 284–6; introduction to 283–4; NYSED case study 287; overview of 376; types of 288–9; *see also* test security
achievement gains 193, 267
achievement tests 3, 5, 9, 10, 13
Additive cheating 239, 240
adequate yearly progress (AYP) standards 16, 232
affected group 165, 169, 174, 175
American Board of Medical Specialties (ABMS) 101, 359–60
analytical reasoning (AR) items 165, 171
answer copying: as approach to cheating 263; cheat sheets and 381; detection of 26–33; index values 126; introduction to 13, 17; simulation studies and 35–6, 38; statistics 47, 48, 49, 54; as threat to test security 289, 293
answer documents: changing answers on 381; erasures on 195, 196; introduction to 8, 9, 11; manipulation of 8, 193, 195; processing of 196–7, 200–1
answer matches by ϕ 130
answer matches by ϕf 130
answer similarity indices 126–7
assessment scores 288, 335, 336, 337, 341
assistance as threat to test security 289, 293
assumptions: effect sizes and 51; violations of 50–3
Atkins v. Virginia 335
ATP security survey 393–4
AUC estimate 36, 38, 40
baseline data: conclusion about 321; creation of 311–12, 319–20; establishing 312–13; for examinees 310, 311, 312; false negatives and positives and 380; index values and 314–19; for licensure assessment 309; purpose of 310; summary about 376; test misconduct and 310–21
Bayesian Hierarchical Linear Model (BHLM): conclusion about 243; introduction to 233; method 233–7; OpenBUGS code and 415–16; results and discussion 237–40; score gain analysis and 241–2; summary about 375
Bayes' rule/theorem: conclusion about 356–7, 392; conditional probability and 349; discussion of 352–6; implementation of 349–50; log odds ratio statistic and 112; method 351–2; posterior distributions and 166; PPoC values 355; *p*-values and 350–1; summary about 376
Belov's method 114, 115
benign erasures 215, 217, 218, 228, 229
between-cohort modeling strategy 249–50
biased item and person parameter estimates 106–7
binomial distribution 27, 31–2, 49–50, 211, 218
bivariate normal distribution 88, 179
bivariate probabilities 61, 65
Bonferroni adjustment 41, 53, 129, 169, 171
brain dump sites 102, 104, 106, 109
certification examinations 3, 151, 152, 160–1, 298
cheating: approaches to 72, 102, 263, 396–7; for benefit of oneself or others 5; countermeasures for 381; defined 4, 72; establishing thresholds for 273–4; extent of 269–73; incidence of 5, 348; inferences about 71; as purposeful 4–5; reasonable transparency and 397–9; scandals related to 284–5; scope of the problem 382–3; as a validity concern 6–11; why it is wrong 5–6; *z*-score and MLR methods and 274–5; *see also* quantitative methods; test fraud
cheating detection: conclusion about 17–18, 390–9; credentialing dataset for 14–15; K-12 education dataset for 15–16; need for 12; PPoC and 237

- cheat sheets 263, 296, 381
 chi-squared distribution 87, 88, 89, 91
 cluster analysis: licensure dataset and 143–7; PPoC and 237; real-data application and 146–7; test collusion and 115, 139–43
 clustering threshold δ 127, 129
 collateral evidence: collection and retention of 360–6; evaluation of 367–8; exam results and 358–60; interviews after exams as 366–7; introduction to 358; test security and 360–1
 collusion *see* test collusion
 combinatorial search 170–1, 175
 compromised items *see* item compromise
 computer-based testing (CBT): DPF and DIF analyses and 160; introduction to 13, 14; item latencies and 318–19; response time and 73
 computerized adaptive testing 13, 76, 101, 124
 confirmatory analysis 53–4, 62
 copying conditions 35, 40, 68
 correct and incorrect responses: identical 27, 30–2, 55–7; introduction to 9–11; response similarity indices and 29; similarity statistics and 48–50
 creative responding 72, 263
 credentialing dataset 14–15, 329, 339, 342
 credentials 3, 17, 119, 359
 criterion-referenced tests 380, 385
 cumulative distribution function (CDF) 197, 198, 199
 cumulative logit model 250–5, 258, 375–6
 cumulative probabilities 251, 342, 343
 Cumulative Sum (CUSUM) statistic 74, 76, 78–9, 81–6, 374
- data: experiments related to 173–4; flagged individuals 108, 109; at item level 331, 332; for pass rates study 268; test fraud 323–44; as a validation tool 17; *see also* simulation design/studies
- data analysis: base rate issues 380; conclusion about 388–9; limitations of comparison and 379; for person-fit indices 78–9, 82–6; for response time models 89–92; score gain 241–3; for standardized tests 234–5; using machine learning algorithm 93–4
- data forensics: as action to dealing with test security 299; conclusion about 391, 393; NYSED case study 288; real-time 394; test fraud and 153; test misconduct and 254
- datasets: credentialing 14–15, 329, 339, 342; erasures analysis and 209–11; introduction to 13; K-12 education 15–16, 226–9; licensure 62–3, 66, 103; NRM and 401–3; pass rates analysis and 275–6; PL model item parameters and 404; simulation methodology 67; through MULTILOG 219
- decision making frameworks 387–8
- deep item pool 298–9
- dependent variables 30, 201–4, 235
- detection threshold 62, 68, 275, 352–3, 355
- Deterministic Gated Item Response Theory Model (DGM) 104–9, 374
- deviations from the mean 204–5, 210, 273
- dichotomous response outcomes 26, 34, 36, 39, 74
- Differential Item Functioning (DIF): conclusion about 161–2, 374; discussion of 160–1; flagging parameters 156–7; introduction to 52; preknowledge and 116, 155–6, 382; results related to 158–9
- Differential Person Functioning (DPF): conclusion about 161–2, 374; discussion of 160–1; flagging parameters 156–7; introduction to 110; preknowledge and 116, 154–5; results related to 157–8
- discrete distributions 168, 170
- distributive justice concept 387
- Divergence Algorithm 169–71, 173–5
- D_0 indices 28, 35–6, 38–9
- Educational Testing Service (ETS) 102, 289, 290
- effective response time (ERT) model 86–7, 89, 91
- effect sizes 9, 51, 384–5
- empirical null distributions 27, 29, 31
- erasure detection index (EDI): description of 215–16; to detect tampering 216, 230; extension of 216; introduction to 215; K-12 dataset and 226–9; summary about 375; Type I errors for 219, 230
- erasures: analysis of 194, 201–9; on answer documents 195, 196; capturing of 15; data frame considerations 200–1; flagged examinees and 210–11; flagging observations and 199–200; fraudulent 216, 218; gain scores and 194–5, 197–9; GEEs and 207–8; history and current status 193; HLM and 206–7; introduction to 9; item difficulty and 315; item sequence and 314; misalignment 217, 218; occurrence of 195; patterns of 10, 11; random 217, 218; simulation studies 216–21, 225–6; string-end 217; summary about 375; types of 195–6; visualization and 208–9; *see also* cheating
- erasure victims 218–26, 228
- error bands 249, 339
- error similarity analysis (ESA) 31, 49
- ESEA flexibility waiver program 285
- evidence: sources of 9, 12; statistical 8, 277, 391–3, 397
- evidentiary standards 368
- exam content: distributing live 104; as a form of cheating 102; harvesting 324; introduction to 101; prior access to 154, 158; *see also* preknowledge
- examinees: aberrant 164–5, 167, 169–74; agreement 360–2, 366; baseline data for 310, 311, 312; collecting personal information about 361–2, 381–2; contaminated or uncontaminated 134, 136–8; detecting items and 116–21; erasures and 210–11; evidentiary standards and 368; flagged 40–2, 106–9, 135, 174, 354, 377; high-ability 35, 38, 40, 104, 109; hit rates 377–8; index values of 312–14, 320; interviewing of 366; latency information 315, 318–19; low ability 17, 38, 39, 40, 113; methods to detect groups of 114–16; misclassification rates 377–8; noncolluding 141, 145; potential collusion among 47, 48, 55; with preknowledge 106, 111, 113; recorded statement of 366; response patterns across 262–3, 365, 366; summary about 376; Type I errors and 129, 132;

- video and audiotaping of 363; *see also* response time (RT) model; test takers
- exams: access to prohibited materials during 364; collecting information before, during and after 361–6; conducting interviews after 366–7; hypothetical investigation after 365–6; results related to 358–60, 362, 365, 368
- exploratory analysis 53
- exposure as threat to test security 289, 293–5
- “extent of cheating” effect 269–73
- factor analysis 115–16
- failing scores 268, 269, 274–6
- false negatives and positives: baseline data and 380; consequences related to 347; introduction to 12; preknowledge detection and 103, 104, 106–8; test tampering and 223–4
- flagged examinees *see* examinees; test takers
- FLOR log odds ratio index 110, 116–19
- forensics *see* data forensics
- fraudulent erasures 216, 218
- $G = 5$ conditions 135, 136, 137
- $G = 10$ conditions 134, 135, 136–41
- gain scores 194–5, 197–9, 241–3
- gaming as threat to test security 289
- Generalized Binomial Test (GBT) 32–3, 35–6, 38–9, 57–9, 215
- generalized estimation equations (GEEs) 207–8
- grand mean 204, 205
- graphs: choice of 325; illustrating pass rates 329–30; sparklines as 331; *see also* visual displays
- growth aberrance 233, 239–43, 415–16
- hierarchical linear modeling (HLM) 206–7, 233, 235, 242
- hierarchical RT model 179, 186–9
- high-stakes testing 70, 286, 308, 376, 393
- hit rates 377–8
- H^T distributions 79, 80, 81
- H^r index 77–8
- human behavior, accountability testing and 284–6
- hypothesis testing 346, 348, 350–3, 356, 357
- identification as threat to test security 289
- incorrect matching answers 50, 56, 59
- incorrect responses *see* correct and incorrect responses
- independent variables: introduction to 30, 31; simulation design and 34, 35, 132; Type I errors and 133, 134, 219
- index values: answer copying 126; baseline data and 312–14; detection threshold and 62; interpretation of 314–19; M4 statistic 67, 68, 127; similarity statistics and 53, 54
- individuals *see* examinees; test takers
- inferences: about cheating 71; about test scores 7; introduction to 6–7; stochastic 233, 236; validity of 152, 177, 189
- investigation as action to dealing with test security 301
- IQ scores 335, 336, 337
- item compromise: conclusion about 122; detection of 102–4, 156–7; high profile cases 101–2; introduction to 101; licensure dataset and 103; methods to identify 110–14; minimizing impact of 153; probability of 118, 119; Rasch model and 112; simulated conditions and 120, 395; summary about 374; testing programs and 102, 105; test scores and 154; Trojan Horse items and 109; *see also* preknowledge
- item degradation 152, 155, 157, 162
- item difficulty estimation 106–7, 110, 155, 156, 315
- item exposure 15, 52, 386
- item harvesting 177, 379, 380, 386, 396
- item latencies 316, 318–19
- item-level aberrance 115, 116
- item-level methods 262–3
- item parameter estimates 86, 116, 131, 147, 148
- item pools 153–7, 292, 298, 324
- item preknowledge: combinatorial search and 170–1, 175; conclusion about 174–5; detection statistics 166–8; Divergence Algorithm and 169–70; introduction to 164; occurrence of 164; problem statement and 164–6; real data experiments and 173–4; simulated data experiments and 171–3; summary about 374; terms of 166; *see also* preknowledge
- item response theory (IRT): collusion detection with 132; conditionally independent 52, 55; introduction to 14; Type I errors and 33; *see also* response time (RT) model
- item statistics 152, 157, 162
- JL method 124–5
- joint distributions 125, 205, 375
- joint flagging strategy 210, 378
- K_1 indices 31, 35, 38–9, 412, 414
- K_2 indices 31, 35, 38–9, 412, 414
- K-12 education dataset 15–16, 226–9
- K index 31, 33, 49, 263
- kitchen-sink approach 390
- Kullback–Leibler Divergence (KLD) 114–15, 168
- licensure dataset: analysis of 62–3; collusion detection and 143–7; Divergence Algorithm and 173–4; item compromise and 103; lognormal RT model and 186
- licensure examinations 3, 13, 14, 34
- linear prediction model 249, 252
- linear regression 205, 206, 210
- logical reasoning (LR) items 165, 171
- logistic regression model 325, 326, 327, 328
- lognormal RT model 179, 181–4, 186
- log odds ratio statistic 112–14, 117–18, 121, 122
- log response time 86–7, 89, 90, 179
- lucky guessing 72, 263
- l_z and l_z^* distributions 79, 80, 81
- l_z and l_z^* indexes 73–6
- M4 similarity statistic: computation of 127–8; index values 67, 68, 127; introduction to 59; licensure dataset and 62–3; presentation of results from 63–6; probability density function and 59–62; summary about 373–4; tail probabilities and 59–62

- machine-learning-based approach 73, 92–5
- Mantel-Haenszel test 154, 155, 156
- market basket analysis 73, 93–5
- Markov Chain Monte Carlo (MCMC) estimation 105, 179, 235–7, 241, 242
- matching answers 49–50, 58–9, 127, 129
- materials as threat to test security 289, 296
- math proficiency categories and levels 16
- measurement error 248, 338, 339, 341, 385
- mental retardation 335, 336
- misalignment erasures and errors 217, 218
- misclassification rates 377–8
- monitoring as action to dealing with test security 298
- Monte Carlo simulation study 33, 233, 241
- moving averages 110–12, 331–2, 374
- multilevel logistic regression (MLR) 265–77, 376
- MULTILOG 34, 219, 227
- multinomial distribution 59, 201
- multiple comparisons 53, 129, 131, 143–5, 257–8
- nearest neighbor clustering method 126, 127
- negative binomial (NB) distributions 199, 201, 208
- New York State Department of Education (NYSED) 287–8
- next generation testing: baseline data and 310–21; introduction to 309; as a mixed format 310; summary about 376; test security and 302–4
- No Child Left Behind Act (NCLB) 4, 232, 284
- nominal response model (NRM) 34–5, 127, 132, 217–19, 401–3
- nominal response outcomes 36, 37, 40
- nonaberrant examinees 165, 168, 171, 172
- nonlinear regression 250, 382
- nonparametric person-fit indexes 77–8
- norm-referenced tests 193, 346, 380, 385
- null condition 67–8, 79, 108–9, 118–19, 348
- null distributions 29, 31, 127, 352, 355
- null hypothesis testing 346, 348, 350–3, 356, 357
- obstruction as threat to test security 289, 296
- ω index: answer copying detection and 32–3; test centers integrity and 40, 41; test takers and 42; Type I error rates and 43
- online testing 289–91, 296–7, 302–3
- OpenBUGS freeware 235, 236, 241, 415–16
- optical scanners 196–7
- paper-based tests 13, 195–6, 289, 291, 302–3
- parametric person-fit indexes 73–4, 78–9
- Partnership for Assessment of Readiness for College and Career (PARCC) 289, 291, 303
- pass rates: baseline 271, 272, 274; comparison of 332; educational dataset and 275–6; estimating 325–8; graph illustrating 329–30; introduction to 262; models for conducting 265–8; population 270, 271, 272, 274, 276; for several countries 331; *see also* simulation design/studies
- peer review process 398–9
- performance levels: categories of 247–51; conclusion about 260; empirical demonstration 255–9; flagging criterion and 254–5; future directions related to 259–60; introduction to 16; model selections considerations 246–50; outcome of 248; predicting 249; scores gain analysis and 195, 197; standard error and 252, 258, 259; standardized residual for 252–4; student demographic variables and 253–4; test data and 255; of test takers 49, 50; *see also* test scores
- performance tasks 286, 289, 291, 293–6
- person-fit indices: for aberrant patterns 73–86; challenge of using 28; conclusion about 43, 95; CUSUM and 78–9, 81–6, 374; data analysis for 78–9, 82–6; for detecting answer copying 26, 28–9; drawback of 76; nonparametric 77–8; parametric 73–4; summary about 27, 374; test takers and 82–4
- person-fit statistics 104, 115
- Poisson distribution/model 27, 31, 201, 202
- policies as action to dealing with test security 299–300
- policy makers 273, 274, 287, 288, 344
- population mean, Z -test for 203–4
- posterior distributions 168, 180, 236, 241, 349
- posterior probability 112, 117, 181, 236, 351
- Posterior Probability of Cheating (PPoC): calculating 351–2; cheating detection and 237; description of 103; growth aberrance and 243; Type I errors and 237, 238, 240; values of 353–5
- power *see* hypothesis testing
- preknowledge: conclusion about 122; detection rates 103–4; DIF and 116, 155–6, 382; discussion of 119–21; DPF and 110, 116, 154–5; examinees with 106, 111, 113; flagging criteria for 121; FLOR log odds ratio index and 110; as a form of cheating 72; identifying individuals benefiting from 104–10; introduction to 13, 14, 17; item compromise and 152–3; removing contamination due to 120; response time models and 119, 120; summary about 374–5; test fraud and 51; test takers and 396
- pressure as threat to test security 289, 297
- pretest items 66, 155–9
- probabilities: estimation issues 52; expression of 341–4; of observed value 61–2; tail values 49–50, 53–4, 58, 61, 67
- probability contour 61, 65, 343
- probability density function 59–62
- procedural justice 387–8
- proctoring/proctors: independent 276, 277; interviewing of 366; as threat to test security 289, 297; training issues 301
- proficiency classifications 16, 264, 265
- proxy schemes 362
- proxy test takers 48, 55, 324
- psychometrics-based approaches: conclusion about 95; description of 73; person-fit indexes 73–86; purpose of 73; response time models 86–92; for testing programs 153

- p*-values: Bayes' rule and 350–1; calculating 180; moving averages of 110–11; null hypothesis 348, 350, 351, 352; test administration and 332
- quantitative methods 8, 11–13
- racketeering convictions 233, 264
- random erasures 217, 218
- random responding 72, 104, 263
- Rasch model: aberrant response patterns and 74; introduction to 14, 15, 16; item compromise and 112; K-12 education dataset and 227
- real data *see* data
- Receiver Operating Characteristic (ROC) curve 28, 35, 39, 119
- recording devices 366
- recovery as action to dealing with test security 301
- regression analysis 30–1, 199, 325
- replications: DGM model and 108; simulation studies and 35–6
- requirements as action to dealing with test security 300
- response behavior 104, 111, 112, 177, 189
- response similarity indices/patterns: conclusion about 43; data manipulation and 405–14; for detecting answer copying 29–33; introduction to 25–6; summary about 27, 373; test centers integrity and 40–1
- response time (RT) model: approaches to 178–9; conclusion about 95, 189–90; data analysis for 89–92; data source and 180–1; description of 86; discussion of 189; effective response time 86–7; flagging items using 184, 186; flagging persons using 180–6; hierarchical framework 87–8, 186–9; introduction to 177–8; preknowledge and 119, 120; research background 178–9; standard deviation and 181; summary about 375
- right-to-wrong (RW) erasures 195, 196
- S_1 indices 31, 35, 38–9, 412, 414
- S_2 indices 32, 35, 38–9, 412, 414
- sampling error 338, 339
- SAS code 255–8
- scanning technology 196–7
- S-Check statistics 49, 55, 57–9
- scores gains 194–5, 197–9, 241–3
- score validation 49, 102, 245, 382
- scoring as action to dealing with test security 300
- security breaches: collusion detection and 143; introduction to 116; investigation of 304–5; NCLB and 284; sources of 393–4; standardized tests and 286; *see also* test security
- selection as threat to test security 289, 297
- signal-to-noise ratio (SNR) 384
- similarity matrix 115, 116
- similarity statistics: analysis of 47–9; Bonferroni equation and 53; correct and incorrect responses and 48–50; discussion of 49–50; index values and 53, 54; limitations of 54–5; modeling the distributions of 55–8; region of permissible values for 56–8; statistical inferences and 49–55; summary about 373–4
- simplicity principle 329–30
- simulated annealing 170–1, 173, 175
- simulation design/studies: answer copying and 35–6, 38; conclusion about 43; datasets and 67; DGM-related 106; discussion of 143; erasures 216–21, 225–6; independent variables and 34, 35, 132; introduction to 12; item compromise and 120; real-data 34–40; realistic scenarios and 121; replications and 35–6; simulated conditions and 395–6; standardized tests 234–5; Type I errors and 131–43; z-score and MLR methods and 268–9
- Smarter Balanced assessments 283, 289, 291
- source-copier approach 47, 67
- sparklines as graphs 331, 332
- specificity, description of 377
- standard deviation (SD): flagged examinees and 108; of log normal distribution 88; of log response time 89; RT model and 181; of WR sums 204, 205
- standard error: of AUC 36; DPF and 155, 156; GEEs and 207; performance levels and 252, 258, 259; standard deviation and 205; WR sums and 204
- standardized log-likelihood of a response vector (ℓ_z) 28–9
- standardized tests 232, 234–5, 260, 284–6
- Standards for Educational and Psychological Testing* 6, 7
- statewide assessments 232, 264, 268, 303, 320
- statistical analysis *see* data analysis
- statistical evidence 8, 277, 391–3, 397
- statistical inferences, similarity statistics and 49–55
- statistic ℓ_z 167, 173
- string-end erasures 217
- student performance 245, 260, 277
- super-clusters 142, 143, 145
- suspicious subgroups 169, 170
- systems as action to dealing with test security 300–1
- tail probabilities 49–50, 53–4, 58, 61, 67
- teacher cheating 124–5, 194, 235
- technology coordinator, role of 303–4
- test administration/administrators: countermeasures for cheating and 381; introduction to 5, 9; online assessments and 302–3; possible preexposure and 316–18; *p*-values and 332; training issues 303
- test centers 40–1, 171–3
- test collusion: cluster analysis and 139–43; conclusion about 147–8; description of 48; detection of 114–16, 126–9, 144, 145, 146, 147; introduction to 124–5; JL method for 124–5; licensure dataset and 143–7; M4 similarity statistic and 59–68; real-data application and 146–7; similarity statistics and 47–58; simulation studies 129–43; teacher cheating and 124–5, 194, 235; *see also* examinees
- test compromise 17, 43, 126, 385
- test data *see* data
- test development process 7, 152, 367, 394
- test fraud: assumptions and 51; Bayesian methods and 346–57; collecting evidence for 361–6; conclusion about 161–2, 368–9; damage caused by 151–2; data forensics and 153; defined 151;

- detection of 62, 262–5, 273–5, 338; hypothetical investigation about 364–6; introduction to 8, 14; preknowledge and 51; school size and 272–3; *see also* cheating; visual displays
- testing behaviors 104, 153, 180
- testing company 182, 184, 188
- testing professionals 71–2, 248, 323
- testing programs: cheating as a concern for 346, 347; considering basics of 380; flagging parameters and 157; introduction to 13–14; item compromise and 102, 105; limitations of comparison and 379; psychometric methods for 153; summary about 376; *see also* test fraud
- test integrity 25, 177, 284, 285
- test misconduct: baseline data and 310–21; conclusion about 321, 391–2; incidents of 308, 309, 313; investigation of 254; item latency and 315–16; possible preexposure and 317–18; techniques to identify 308; test scores and 245, 246, 255; *see also* examinees
- test performance: cohort modeling approaches 249–50; cumulative logit model for 250–5, 258; identifying unusual changes in 246, 250–5; introduction to 3, 4; model selections considerations 246–50; outcome measure of interest and 247–8; teachers' job performance and 284; uncertainty in predicted outcomes and 248–9; *see also* performance levels; test scores
- test responses 48, 51, 52
- test results, test sponsor and 358–60
- test scores: evaluating accuracy of 71; inferences about 7; integrity issues 42–3; interpretation of 18; item compromise and 154; manipulating 240, 245, 246; negative correlations with 52; test misconduct and 245, 246, 255; test-taking time and 333–4; validity of 7–8, 12–13, 70, 308, 382; *see also* performance levels; test performance
- test security: ABMS's view on 359–60; case study 287–8; collateral evidence and 360–1; conclusion about 304–5, 389, 392–3; confirmatory analysis and 54; cost of maintaining 286–7; fraud detection and 264–5; high profile cases 101–2; introduction to 3, 4; need for improvements in 285–6; next generation testing and 302–4; policy issues 277; policy makers and 273, 274, 287, 288, 344; test development process and 394; threats to 288–302, 381, 391; various actions in dealing with 298–301, 302; vertical integration of 360; visual displays and 324; *see also* response time (RT) model; test collusion
- test sponsor: challenges for 383–8; collusion detection and 144, 145, 146, 147; communicating results to 386–7; evidentiary standards and 368; examinee agreement and 362; test results and 358–60; test security and 360–1; *see also* test fraud
- test takers: chi-squared values 87, 88, 89, 91; credentialing dataset and 14; features associated with 94; introduction to 4, 5, 9; irregular behavior by 177, 178; M4 similarity analysis 63–6; mean exam score and 339–40; nonindependent 50, 55; pass rates for 332; performance levels of 49, 50; person-fit indices and 82–4; preknowledge and 396; proxy 48, 55, 324; response behavior by 111, 189; RT patterns for 184, 186; similarity statistics and 56–7; test centers integrity and 40; *see also* cheating
- test tampering: conclusion about 229–30; detection of 216–21, 225–6; introduction to 13; tampered classes and 218–19, 221–6; tampered items and 218–23, 225; *see also* erasures
- TE sums 203, 205, 206
- 3D Algorithm 169, 175
- training as action to dealing with test security 301
- trinomial distribution 50, 55, 59–61, 127–8
- Trojan Horse items 109–10, 153
- t*-test 30, 155, 156
- two-proportion z-score approach 265–77, 376
- Type I errors: of all indices 36, 37; clusters containing 132, 134, 140, 142–4; collusion detection and 131–43; for EDI 219, 230; examinees and 129, 132; inflation in 396–7; introduction to 33; for M4 similarity 66–8; PPoC and 237, 238, 240; test centers integrity and 40, 171–3
- Type II errors 54, 66–8
- U3* distributions 79, 80, 81
- U3* index 77
- uncertainty indications, visual displays and 338–41
- validity: defined 6; of test scores 7–8, 12–13, 70, 308; triangle 71–2; *see also* cheating
- Vela v. Nebraska* 335
- visibility as action to dealing with test security 301
- visual displays: audience as primary consideration 336–8; choice of 324–5; conclusion about 344; features of 323–38; integrating numbers and figures for 330–3; interpretation of 331; introduction to 323; probabilities depiction and 341–4; simplicity principle for 329–30; summary about 376; test security and 324; uncertainty indications and 338–41
- visualization, erasures and 208–9
- VM index 27, 33
- Winsteps software 154, 155, 156
- within-cohort modeling 249–50
- wrong-to-right (WR) erasures 11, 194–6, 214, 228–9, 321
- wrong-to-wrong (WW) erasures 195, 196
- WR sums 201–5, 211
- WR/TE ratio 202–3
- z-score method 265–77
- Z-test 203–5, 209–10