RELATIONSHIP BETWEEN TEST SECURITY POLICIES AND TEST SCORE MANIPULATIONS

by

Bradley Adam Thiessen

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Psychological and Quantitative Foundations
(Educational Measurement and Statistics)
in the Graduate College of
The University of Iowa

December 2008

Thesis Supervisor: Associate Professor Timothy Ansley

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Bradley Adam Thiessen

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Psychological and Quantitative Foundations (Educational
Measurement and Statistics) at the December 2008 graduation.

Thesis Committee: Timothy Ansley, Thesis Supervisor

Stephen Dunbar

Lelia Helms

Andrew Ho

Michael Kolen

ACKNOWLEDGMENTS

I thank Andrew Ho, Tracey Magda, and Katherine Furgol (a.k.a. *The Team*) for their invaluable assistance in collecting and verifying state test data. Their help turned a seemingly impossible task into a mildly painful experience.

I also thank my advisor, Dr. Ansley, for commenting on early drafts of this paper and allowing me the freedom to work through problems independently. I owe my interest in aberrant response detection to a project I completed for Dr. Ansley's IRT class. When he asked me why I chose aberrant response detection as my project topic, I remember replying, "Mostly because it was first on the list."

I also thank Dr. Ansley for his "even statistics has its moments" joke that I have since stolen and used in all my statistics classes.

I am grateful for the helpful suggestions made by the members of my committee: Dr. Dunbar, Dr. Helms, Dr. Ho, and Dr. Kolen. Their comments *may* have caused the quality of my paper to increase, although I am still investigating plausible rival hypotheses.

In addition to the committee members listed above, I thank the other great teachers who have gotten me to this point in my academic career. I thank Dr. Frisbie and Dr. Qualls for proving that learning about assessment can be fun and for providing me with a great start to my graduate school career. I thank Dr. Brennan for challenging me and bridging the gap between measurement and statistics that existed in my mind.

I thank my parents and brothers who, upon discovering I was writing my acknowledgment, suddenly thought of many ways in which they helped me. Finally, I thank Eva Horváthová for forcing me to take breaks from writing and pretending to be interested when I would talk about my progress.

ABSTRACT

Published news reports, surveys, and previous research indicate that some educators have responded to the current high-stakes testing environment by manipulating test scores. This study first classifies the methods educators have used to manipulate test scores and then explores possible reasons why educators manipulate test scores. Next, possible methods to deter manipulations are explored, including the development and implementation of high quality test security policies. It is found that effective test security policies have four main components. These components are used to evaluate the test security policies currently implemented by all 50 states. These evaluations result in recommendations for improvements in test security policies.

As a possible indicator of test score manipulations, comparisons are made between state test score trends and NAEP trends in reading and mathematics in grades 4 and 8. Due to technical limitations in comparing simple changes in proficiency rates, a scale-invariant framework based on P-P plots is used to estimate state-NAEP score trend discrepancy effect sizes. These effect sizes show that state trends were significantly larger than NAEP trends from 2003-05, 2005-07, and 2003-07.

After discussing plausible hypotheses for these score trend discrepancies, this dissertation examines the relationship between the quality of state test security policies and these scale-invariant score trend discrepancy effect sizes. No significant relationships were found for any of the time periods, subjects, or grade levels. Longitudinal analyses and group mean comparisons were also unable to find any significant relationships between test security policy quality and score trend discrepancies.

TABLE OF CONTENTS

LIST OF TABLES vi

LIST OF FIGURES ix

CHAPTER 1: INTRODUCTION 1

Background 1
Statement of the Problem 3
Purpose and Research Questions 4
Significance of the Study 5

CHAPTER 2: LITERATURE REVIEW 8

Definition and Methods of Manipulation 8
Prevalence of Test Score Manipulations 17
Prevalence Estimates Based on Teacher Surveys 21
Prevalence Estimates Based on State Surveys 26
Prevalence Estimates Based on Direct Observation 28
Prevalence Estimates Based on Statistical Detection 30
Prevalence Estimates Based on Targeted Research 33
Why Do Educators Manipulate Test Scores? 38
Educators Are Former Students 38
Pressure From State Accountability Systems 40
Educators Unaware of Manipulations & Their Impact 43
Lack of Oversight and Policies 50
How to Prevent Manipulations: Evaluation of State Policies 53
Test Security Policy Content: Evaluative Framework 56
Relationship Between Test Security Policies & Trend Discrepancies 66
Single-Year Comparisons of State and NAEP Results 69
Trend Comparisons of State and NAEP Results 75
Scale-Invariant Trend Comparison Methods 80
State and NAEP Trend Discrepancies: Plausible Rival Hypotheses 81
Summary 84

CHAPTER 3: METHODOLOGY 85

Independent Variable: Test Security Policy Quality 86
Data Collection and Verification 87
Sampling 87
Analysis 91
Technical Quality of Policy Evaluation Data 93
Assumptions and Limits 94
Dependent Variable: Scale-Invariant State/NAEP Trend Discrepancies 95
Data Collection and Verification 96
Sampling 100
NAEP Data Collection 102
Analysis 102
Reporting 115
Assumptions and Limitations 116

Analysis 118
Scope 120
Assumptions, Limitations, and Confounding Variables 120
Summary 121

CHAPTER 4: RESULTS AND ANALYSIS 123

Quality of State Test Security Policies 123
Policy Evaluation Component Scores 126
State Policy: Dichotomizations 142
Changes in Policies Over Time 153
Summary: Test Security Policy Quality 154
State-NAEP Trend Discrepancy Estimates 154
Discrepancy Estimates for 2003-05, 2005-07, and 2003-07 155
Discrepancy Estimates by Grade and Subject 160
Summary 162
V Estimates Compared to Other Measures of Trends 162
Choice of 3 Cut-Score Minimum 166
Summary: State-NAEP Trend Discrepancies 169
Relationship Between Security Policy Quality & Trend Discrepancies 170
Regression Analyses 176
Reasons For Lack of Significant Relationships 180
Categorical Analyses 180
Longitudinal Analyses 189
Summary 190

CHAPTER 5: SUMMARY, DISCUSSION, AND RECOMMENDATIONS 192

Summary 192
Discussion 193
Limitations 197
Recommendations 199

APPENDIX A: PUBLISHED NEWS SUMMARIES 201

APPENDIX B: STATISTICAL DETECTION INDICES 214

Early Developments 214
From Empirical to Chance Models 215
Incorporating More Information 217
Controlling for False Positives 220
Incorporating Item Response Theory 221
Person-Fit and Aberrant Response Indices 223
Aberrant Response Indices to Detect Examinee Cheating 225
Adjacent Seating Methods 228
Methods to Detect Educator Cheating 229
Index #1: Unusual Test Score Fluctuations 231
Index #2: Unexpected Patterns in Student Answers 233
Combining Indices to Detect Cheating Classrooms 237

APPENDIX C: NATIONAL TESTING CODES AND STANDARDS 238

REFERENCES 241

LIST OF TABLES

Table 2.1	Taxonomy of Manipulations in Published News Reports (1994-2007)	11
Table 2.2	Estimated Prevalence of Manipulations	19
Table 2.3	Percentage of pre-service teachers rating each activity as appropriate	46
Table 2.4	Median percent of Iowa teachers within schools rating each activity as ethical or unethical	46
Table 2.5	Teacher and testing specialist mean appropriateness ratings	47
Table 2.6	Examples of appropriate and inappropriate test preparation practices	48
Table 2.7	Guidelines and criteria to select appropriate test preparation activities	49
Table 2.8	State and school district test security policies	51
Table 2.9	Content Recommendations for Test Security Policies	60
Table 2.10	FOIL Framework for Evaluating Test Security Policy Content	64
Table 2.11	Results from 2005 state and NAEP tests of 8th grade mathematics	70
Table 2.12	Discrepant trends in 8 grade reading achievement (2003-2005)	79
Table 2.13	Discrepant trends in 8 grade reading achievement (2003-2005)	79
Table 3.1	Scale-Invariant Trend Discrepancy Data	96
Table 3.2	General test information collected from each state DOE website	98
Table 3.3	Example of test score data to be collected	99
Table 3.4	Example of collected test data to interpolate P-P curves	110
Table 3.5	Data included in the analysis	113
Table 3.6	Data excluded from the analysis	114
Table 4.1	Policy evaluation scores for each state	124
Table 4.2	Spearman rank-order correlations between policy component scores	125

Table 4.3	Summary of policy evaluation ratings	127
Table 4.4	Formalize subcomponent distributions	130
Table 4.5	Oversee subcomponent distributions	134
Table 4.6	Inform subcomponent distributions	137
Table 4.7	Limit subcomponent distributions	140
Table 4.8	Policy Categorizations	143
Table 4.9	Policy evaluation composite scores by category	144
Table 4.10	Test security incidents and responses in Texas from 2005-07	148
Table 4.11	Changes in state policy content or procedures from 2005-07	153
Table 4.12	Average state-NAEP discrepancy estimates	156
Table 4.13	Rank-order correlations	173
Table 4.14	Rank-order correlations	174
Table 4.15	Regression analyses and coefficients of determination	177
Table 4.16	Standardized regression coefficients	178
Table 4.17	Summary of trend discrepancy estimates by categorizations	181
Table 4.18	Consistency of group differences across time periods	183
Table 4.19	Average trend discrepancies for state- and district-level policies	185
Table 4.20	Groups within treatments analysis for state vs. district level policies	186
Table 4.21	Unweighted mean state-level trend discrepancies for each categorization	187
Table 4.22	Groups within treatments analysis for state vs. district level policies	188
Table 4.23	Groups within treatments analysis for independent monitoring policies	188
Table 4.24	Groups within treatments analysis for policy tone	189

Table 4.25 Longitudinal Analyses 190

Table B.1 Results from 2005 state and NAEP tests of 8th grade mathematics 229

Table B.2 Example data for Index #1 233

LIST OF FIGURES

Figure 2.1	The number of published news reports on manipulations from 1995 – 2007	18
Figure 2.2	The number of published news reports from each state (1995 – 2007)	18
Figure 2.3	Score trends in 8 grade mathematics measured by the MSPAP, TAAS and NAEP	77
Figure 3.1	Evaluation form for state test security policies	88
Figure 3.2	Example of composite scores calculated for each state	93
Figure 3.3	Example of data collected from CCSSO.org website	97
Figure 3.4	Simulated test score distributions to illustrate pliability of PPS-based trends.	104
Figure 3.5	CDFs from a test administered at Time 1 and Time 2 with cut-scores of 500, 700, and 800	105
Figure 3.6	P-P plot from the simulated data displayed in Figures 3.3 and 3.4	107
Figure 3.7	Smoothed (interpolated) P-P plots for the example data in Table 3.4.	111
Figure 3.8	Scatterplot to display scale-invariant trend effect sizes	115
Figure 4.1	Scatterplot to display scale-invariant trend effect sizes	126
Figure 4.2	Distributions of security policy evaluation ratings	128
Figure 4.3	State-NAEP discrepancies for 2003-05, 2005-07, 2003-07	158
Figure 4.4	Relationship between 2003-05 and 2005-07 State-NAEP trend discrepancies.	159
Figure 4.5	Summary of state-NAEP trend discrepancies by subject and grade	161
Figure 4.6	Relationship between *d* and *V* effect size estimates for state trends	164
Figure 4.7	Relationship between *d* and *V* effect size estimates	164
Figure 4.8	Relationship between proficiency trends and *V* discrepancies	165

Figure 4.9	P-P plots and *V* estimates after eliminating one cut-score	168
Figure 4.10	Average state-NAEP discrepancies for each state (2003-05, 2005-07)	170
Figure 4.11	Scatterplots of policy quality composite scores and discrepancy estimates	171
Figure B1	Aberrant Response Detection Methods and Indices	224

CHAPTER 1: INTRODUCTION

> Preventing cheating by those who give tests is a particularly underresearched topic. It is ironic that much attention has been given to preventing cheating by individual students – behavior that can cause a single score to be of questionable value – and so little attention has been paid to cheating by those who give tests, which can invalidate the scores of entire groups of students.
>
> Gregory J. Cizek, *Cheating on Tests: How to Do It, Detect It, and Prevent It*

Background

In an effort to increase the academic achievement of all students and confront the "soft bigotry of low expectations" (Bush, 2000), the *No Child Left Behind* (NCLB) Act was passed into law on January 8, 2002. Like the *Improving America's Schools Act* (IASA) signed in 1994, NCLB required all states to develop content and performance standards; implement assessment systems to track student performance against those standards; and create adequate yearly progress (AYP) goals to ensure all students reach a proficient level of achievement (IASA, 1994; NCLB, 2001). Believing the IASA was ineffective in improving student achievement due to its status as "an undertaking without consequences," (Rotherham, 1999) NCLB granted the federal government the authority to impose sanctions upon schools, school districts, and states failing to meet AYP goals. The sanctions were intended to provide an incentive for educators to improve the quality of education provided to students and, ultimately, to improve student achievement.

Incentive theory predicts that by tying the threat of sanctions to assessment results, NCLB would motivate educators to increase test scores by implementing more effective instructional programs and, as a result, student achievement will increase (Laffont & Martimort, 2001; Jacob, 2007). Researchers have found evidence of this effect, finding that students in states with accountability systems achieve significantly higher gains on the NAEP (National Assessment of Educational Progress) math test than students in states having no sanctions tied to assessment results (Carnoy & Loeb, 2002; Hanushek & Raymond, 2004ab, 2006). These researchers conclude that "the introduction

of consequential accountability systems has a clearly beneficial impact on overall performance" (Hanushek & Raymond, 2004a, p. 32) for students of all racial groups even after controlling for test participation rates and state characteristics (Carnoy & Loeb, 2002, p. 318). Other researchers have also found this beneficial impact of NCLB on student achievement, concluding that the strength of a state's accountability system "is indeed an important predictor of student performance at all points on the distribution curve, and especially so for students at the basic level" (Loeb & Strunk, 2005, p. 23) and that "students perform better than expected when their test score is particularly important for their schools' accountability rating" (Reback, 2007, p. 1).

Although these studies found positive effects of accountability systems on student achievement, other evidence suggests educators are "gaming the system" to increase test scores without a corresponding increase in student achievement. Some educators game the system by manipulating the teaching process or their teaching philosophies. They do this by narrowing the curriculum to primarily teach content found on the test, using actual test items as practice for the test, focusing instructional resources only on the students most likely to improve the school's test scores, spending an inordinate amount of time on test preparation, or by bribing students for higher test scores (Neal & Schazenbach, 2007; Jacob, 2005; Nichols & Berliner, 2004). Other educators have been found to manipulate the test administration by giving students hints on test questions, changing student answers, reading questions aloud to students, or by providing students extra time to complete the test (Cullen & Reback, 2006; Figlio & Getzler, 2002; Jacob & Levitt, 2003; Nichols & Berliner, 2005). Others attempt to game the system by excluding students from testing, inappropriately classifying examinees as disabled, or by using other methods to manipulate the examinee pool (Cullen & Reback, 2006; Figlio, 2005; Figlio & Getzler, 2002). Still others game the system by manipulating student test scores or by lowering proficiency standards (King, 2007; Nichols & Berliner, 2005).

Statement of the Problem

Reports of these manipulations cast doubt on educators and the inferences made from test scores. Tests questions are intended to represent a sample of a larger domain of interest and test scores are intended to represent examinees' performance in this domain (Haladyna & Downing, 2004). Manipulations that increase test scores without correspondingly increasing examinee performance in the larger domain destroy the validity of inferences made from the test scores. Therefore, these attempts to game the system negatively impact any decisions made on the basis of test scores, including the evaluation of instructional programs and the allocation of educational resources.

To protect the validity of inferences made from test scores, several methods have been used to deter educators from gaming the system. Professional organizations have developed ethical codes, guidelines, and standards to inform educators of the negative impact of test manipulations (AERA, APA, & NCME, 1999; JCTP, 2004; NEA, 1990; Schmeiser et al., 1995), but research suggests that educators are still unaware of what behaviors are appropriate or inappropriate (Kher-Durlabhji & Lacina-Gifford, 1992; Lai & Waltman, 2007; Moore, 1994). Several state departments of education and test publishers employ statistical methods in an attempt to detect educators who game the system, but these methods can only detect the most blatant manipulations, and research has shown their limited statistical power (Chason & Maller, 1996; Iwamoto, Nungester, & Luecht, 1996; Impara, Kingsbury, Maynes, & Fitzgerald, 2005). Some states have tried to deter educators by outlining harsh sanctions for anyone caught cheating, but research has shown that test manipulations are not frequently reported (Gay, 1990) and that many states do not follow through with the sanctions (Mehrens, Phillips, & Schram, 1993; Sorensen, 2006).

One promising way in which state officials have attempted to deter educators from gaming the system is through the development, implementation, and dissemination of comprehensive test security policies to both discourage test manipulations and

encourage ethical behavior. While little research has been conducted to determine the effectiveness of these policies on deterring educators from gaming the system (Cizek, 1999), similar policies have been shown to be effective in reducing student cheating on tests at the postsecondary level (McCabe & Trevino, 1993, 2002). If found to be effective in deterring educators from manipulating test scores, states could develop and implement test security policies to ensure inferences made from test scores are valid.

Purpose and Research Questions

This study will first document and classify methods educators use to manipulate test scores. News reports, surveys, observational studies, and statistical detection studies will be used to develop a taxonomy of manipulations and estimate the prevalence of various manipulation methods. This study will then explore possible reasons why educators manipulate test scores.

With information about how and why educators manipulate test scores, this study will then document and classify policies and procedures used by states in an attempt to deter educators from manipulating test scores. A framework will then be developed to evaluate the quality of these state test security policies. This framework will be based on the work of Cizek (1999), McCabe and Trevino (1993, 2002) in designing honor codes and test security policies to deter student cheating on tests.

Finally, this study will attempt to determine if a relationship exists between the strength of a state's test security policy and the discrepancy in score trends between the state tests and an audit test. The logic is that if manipulations increase test scores without a corresponding increase in student achievement, then discrepancies between score trends on the state test and another (audit) test of the same domain could possibly provide evidence of test score manipulations. A scale-invariant framework will be used to estimate trends on high-stakes reading and mathematics tests used in state accountability

systems and trends on the relatively low-stakes NAEP reading and mathematics tests for grades 4 and 8.

While the discrepancies between state test and NAEP score trends could possibly provide evidence of test score manipulations, it is important to note that any discrepancies could be explained by any combination of other plausible rival hypotheses. This study will discuss other possible explanations for the discrepancies between state and NAEP score trends, including possible differences in test content and administration; examinee pool and examinee motivation; and the strength of a state's testing program.

In summary, this study will attempt to address the following research questions:

1. What kinds of manipulations do educators use to increase test scores? Why do educators manipulate test scores? What is the estimated prevalence of each type of manipulation?
2. What test security policies and practices do states implement in an attempt to deter educators from manipulating test scores? What is the quality of each state's test security policy?
3. What is the relationship between the quality of a state's test security policy and any discrepancies between score trends on state and NAEP tests? Which aspects of a state's test security policy seem to have the strongest relationship with score trend discrepancies? What are some potential explanations for the discrepancies between state test and NAEP score trends?

Significance of the Study

This study will synthesize research on inappropriate testing practices to provide a taxonomy of methods used by educators to game the system. Results from surveys (Gay, 1990; Hall & Kleine, 1992; Lai & Waltman, 2007; Mehrens, Phillips, & Schram, 1993; Nolen, Haladyna, & Haas, 1992; Pedulla, et al., 2003; Shepard & Dougherty, 1991; Sorenson, 2006), direct observations (Horne & Gary, 1981; White, Taylor, Carcelli, &

Eldred, 1981; Wodtke, Harper, Schommer, & Brunelli, 1989), test preparation research (Moore, 1994; Popham, 1991), and statistical analyses (Perlman, 1985; Jacob & Levitt, 2004; Wesolowsky, 2000) will be combined with published news reports (Nichols & Berliner, 2004; Thiessen, 2007) to estimate the prevalence of each method. This extends the work of Haladyna, Nolen, and Haas (1991) to determine sources of test score pollution and the work of Jacob and Levitt (2004), Cizek (1999), and Wesolowsky (2000) in determining the prevalence of educator cheating on achievement tests. This study will also synthesize the above research in an attempt to explain why educators manipulate test scores.

This study will extend the work of McCabe and Trevino (1993, 2002) and Cizek (1999) in examining test security practices and honor codes in educational organizations. This study will synthesize professional codes and state policies designed to deter educators from manipulating test scores and develop a framework to evaluate test security policies.

Instead of analyzing examinee- and classroom-level data to estimate the prevalence of test score manipulations (Jacob & Levitt, 2004; Wesolowsky, 2000), this study will attempt to determine if a relationship exists at the state-level between the strength of a state's test security policy and the discrepancy in score trends on two tests of the same domain. This will extend the research of Klein, Hamilton, McCaffrey, and Stecher (2000); Linn, Baker, and Betebenner (2002); Peterson & Hess (2005, 2006); Koretz (2005), and Wei, Shen, Lukoff, Ho, & Haertel (2006) into the discrepancy between state and NAEP test scores.

Rather than using scale-dependent methods of comparing proficiency rates (Education Week, 2006; Lee, 2006; Thomas B. Fordham Foundation, 2005), or mapping state cut scores onto the NAEP score scale (Braun & Qian, 2007; Jacob, 2007; Linn, 2005; McLaughlin et al., 2002), this study will promote the use of scale-invariant methods to compare discrepancies in score trends between state and NAEP tests. This

will extend the work of Ho (2005, 2007) and help support or refute conclusions regarding the discrepancies between state and NAEP trends. By discussing possible explanations for these discrepancies, this study will extend the work of Hill (1998), Koretz (1999, 2005), and Koretz, McCaffrey, and Hamilton (2001).

This study will contribute to an understanding of test score manipulations and test security policies. Results of this research could be used by states in developing, improving, or auditing their current test security policies. It could also provide information that could be used in professional development to train teachers in appropriate test preparation and administration activities. This study could also contribute to the debate over the effectiveness of accountability systems and sanctions in improving student achievement.

CHAPTER 2: LITERATURE REVIEW

The review of literature is divided into five sections. The first section defines test score manipulations and provides examples of the four main ways in which educators manipulate test scores. The second section synthesizes estimates of the prevalence of each manipulation method to demonstrate the problem of test score manipulation. The third section then summarizes research into why educators would manipulate test scores to provide a better understanding of the problem. The fourth section then examines methods used to prevent test score manipulations and provides a framework for evaluating the quality of state test security policies and practices. The fifth section then summarizes research into discrepancies between state test and NAEP results, including a discussion of possible explanations for discrepancies.

Definition and Methods of Manipulation

Inferences made about a student's performance on an underlying construct, such as reading comprehension or mathematics problem solving, are made on the basis of observed test scores. It is normally assumed that an increase in test scores reflects an increase in student performance on the underlying construct. This is not always the case, however, as educators can implement practices to artificially increase test scores. The term *manipulation* will be used to describe any practice used by educators to increase student test scores without an equal, corresponding increase in student performance on the underlying construct.

The definition of *manipulation* is influenced by related concepts in the literature. Messick (1984) used the more general term *construct-irrelevant variance* to refer to the influence on test scores of any factor unrelated to the underlying construct (p. 216). Haladyna, Haas, & Nolen (1990) defined a similar concept of *test score pollution* to refer to situations in which test scores are distorted by factors unrelated to the construct being

tested (p. 9). The term *manipulation* is more specific than these other terms in that it refers only to test score distortions caused by educators' practices.

The term manipulation is also defined to be as general and value-neutral as possible. Research into educator test preparation practices use the terms *inappropriate* (Moore, 1994; Popham, 1991) or *unethical* (Lai & Waltman, 2007) to refer to practices that may distort test scores. Likewise, research into institutional cheating defines *cheating* as "a deception used to misrepresent student achievement" (Haladyna & Downing, 2004, p. 25). These terms imply malice on the part of educators. A manipulation is defined as any practice that distorts test scores, whether the practice was implemented maliciously or with the best of intentions. Also, while cheating and inappropriate or unethical test preparation practices are specific types of manipulations, the term manipulation refers to a broader collection of practices used by educators to inflate test scores.

In an attempt to develop a comprehensive list of manipulations used by educators, published news reports of alleged educator misconduct were collected using the LexisNexis® database and subscriptions to the Google® news alerts system. These reports, summarized in Appendix A, were combined with reports summarized by Nichols and Berliner (2004) in their critique of accountability systems in public education, *The Inevitable Corruption of Indicators and Educators Through High-Stakes Testing*, to yield a total of 186 published news reports of incidents from 1994 through 2007 in which educators in American public schools manipulated test scores. These reports of incidents were then combined with results from related research (discussed in the next section) to develop a list of 35 manipulations used by educators.

In order to better understand the methods educators use to manipulate test scores, a taxonomy was developed to categorize the 35 manipulations. Table 2.1 displays this taxonomy along with the number of published incidents for each manipulation. Under this taxonomy, manipulations are classified into one of four categories: manipulations of

the teaching process or philosophy, manipulations of the examinee pool, manipulations of test administration, or manipulations of score reports or scoring standards.

The first, and broadest, category of methods used by educators to manipulate test scores is through manipulations of the process or philosophy of teaching. These manipulations take place before the test is administered to students. These methods include questionable test preparation practices such as practicing with items identical or similar to those on the actual test, practicing with items from previous years' tests, purchasing commercial test preparation packages, and teaching test-taking skills. Practicing with items identical or similar to those on the test fits the definition of manipulation, because students who practice with these items will earn test scores that may not accurately represent their achievement in the domain of interest (Moore, 1994; Popham, 1991). The use of commercial test preparation packages and the teaching of test-taking skills also fit within the definition of manipulation, because they serve to increase test scores on a specific test or test format without necessarily increasing student achievement in the underlying domain (Lai & Waltman, 2007). Finally, educators who focus instructional resources on specific students who have the best chance at improving test scores at a classroom, school, district, or state level at the expense of other students are also manipulating the teaching process to increase test scores without a corresponding increase in overall student achievement (Neal & Schanzenbach, 2007).

Table 2.1 Taxonomy of Manipulations in Published News Reports (1994 - 2007)

	News Reports 1994-2000	New Reports 2001-2007
Manipulate Teaching Philosophy or Process (before test administration)	**18**	**67**
Examining the test or making copies prior to test administration (piracy)	8	35
Practicing with items identical or similar to the test	7	24
Practice with last year's (alternate form) test items	---	1
Practice with items of the same format as the test	---	---
Use commercial test preparation packages	---	---
Teaching test-taking skills; test-wiseness	---	---
Teaching content from specific test items	3	5
Focusing resources on students who are closest to proficiency	---	2
Primarily teaching content found on the test	---	---
Manipulate Examinee Pool (before or during test administration)	**6**	**18**
Excluding students from testing (encouraging drop outs; suspending students)	5	13
Bribing or paying students to increase test scores	1	1
Having high-scoring students take the test multiple times	---	2
Providing inappropriate special education placement	---	1
Increasing the caloric content of school meals to increase scores	---	1
Manipulate Test Administration (during test administration)	**25**	**127**
Altering a student's answer sheet (changing student answers)	8	33
Sanitizing answer sheets (cleaning answer sheets before scoring)	---	1
Not following test administration procedures exactly	---	---
Giving students answers	7	16
Checking student answers and/or pointing out incorrect answers	5	25
Giving students hints on test items (verbal or nonverbal)	2	19
Rephrasing test items for students	---	2
Allowing students to work together during testing	---	1
Ignoring students who are cheating	---	2
Giving students additional examples	1	7
Providing students extra time	1	7
Reading items that are supposed to be read by students	---	5
Answering questions about test content	---	1
Providing students with reference materials or tools during testing	1	7
Instructing students to fill-in specifics for unanswered items	---	1
Providing inappropriate accommodations to students	---	---
Review skills that will be on tomorrow's test	---	---
Manipulate Score Reports or Standards (after test administration)		
Removing or changing student test scores on official records	1	10
Moving or providing students with false IDs so scores won't count	1	1
Misrepresenting data	---	---
Changing the criteria for proficiency or making the test easier	---	---

The following three news report summaries show recent incidents in which educators have been caught allegedly manipulating the teaching process:

> *The Dallas Morning News,* July 13, 2007 (Benton, 2007b).
> A state investigation finds that David Tamez, an elementary school teacher in Amarillo, Texas, leaked the fourth-grade writing test prompt on the spring TAKS writing test to colleagues before the test administration. Tamez reportedly leaked the test information because he believed educators in other districts were doing it as well. The teacher obtained the test information by volunteering to serve on the committee that selects questions for the final form of the TAKS. He alleges that committee members "regularly smuggle out secret TAKS information to share in the home districts." Another teacher interviewed by investigators signed a statement indicating that Tamez "bragged that the source of his insider test information was… a person he had sex with who works for a company that helps build the TAKS." The Amarillo Independent School District concluded that the teacher obtained the information from an unidentified employee at Pearson Educational Measurement. Tamez resigned from his position, but will retain his teaching certificate if he cooperates with the investigation.
>
> *Dayton Daily News*, February 4, 2007 (Elliott, 2007)
> A newspaper investigation found that students at City Day Elementary School in Dayton, Ohio were given 44 practice questions that were identical or "substantially the same" as questions from the actual state exam. In some questions on the practice test, only names or small details were changed from the real test questions. The investigation was launched due to the suspiciously large amount of improvement shown by the school. In 2005, no sixth grade student in the school passed the math subtest of the Ohio Achievement Test. One year later, 100% of these students (now in 7th grade) passed the math test.
>
> *The Columbus Dispatch*, October 22, 2006 (Richards, 2006a).
> Of the 28 Ohio school districts analyzed by The Columbus Dispatch, 15 had instances of educators cheating on standardized tests. Barbara Oaks, a teacher in the Coventry district, looked through the test and wrote out a geometry problem she thought her students would have trouble with. Winifred Shima, a teacher from the Parma district, used a copy of the test to create a study guide for students that included 45 of the 46 actual test questions. Brian Wirick (East Knox) and Heather Buchanan (Wapakoneta) both used the test to create study guides for students. Judy Wray, a veteran teacher in Marietta, made copies of the actual state test to help students prepare. Wray is reported to have said that teachers cheat more than administrators know.

A second way in which educators manipulate test scores is by manipulating the examinee pool. Rather than increasing the achievement of all students, educators using this method attempt to exclude low-ability students from taking the test or convince high-ability students to take the test multiple times. To exclude low-ability students from testing, educators have resorted to suspending students during the test administration period (Figlio, 2005) or inappropriately classifying students as being disabled (Cullen & Reback, 2006; Figlio & Getzler, 2002). These actions would increase test scores at a classroom, school, district, or state level without actually increasing the achievement of all students, so they fit the definition of manipulations.

The following three news report summaries show recent incidents in which educators have been caught allegedly manipulating the examinee pool:

> *San Francisco Chronicle,* July 16, 2007 (Asimov, 2007ab).
> The California Department of Education concludes that for the second consecutive year, educators at University Preparatory Charter High School in San Francisco interfered with state-mandated testing. State investigators seized illegal copies of the 2005 form of the test that was used to prepare students for the exams. Eight former teachers at the school assert the existence of a culture of cheating at the school. According to those former teachers, student grades are frequently falsified and low-scoring students are excluded from state-mandated testing. Last year, the state found that hundreds of answers on the ninth-grade English and math tests had been changed from wrong to right. A counselor from Oakland's Skyline High school reports that a student earning D's and F's transferred to University Preparatory Charter High School and received A's and B's while taking 16 classes in a single semester. When the student returned to Skyline High, he once again earned D's and F's. Last year, investigators concluded that educators at the school changed hundreds of test answers before they were sent for scoring. Former testing coordinator Mike Schwartz is suing school founder and director Isaac Haqq for breach of contract, claiming Haqq was responsible for the altered answer sheets.

> *Brevard School District,* June 30, 2006 (Brevard SD, 2006)
> Lori Backus, principal of Cocoa High School in Brevard, FL is accused of moving at least 54 9th and 10th grade special needs students into 11th grade so that their FCAT scores would not count towards the school's grade (assigned by the state) in 2005 and 2006. As a result of an investigation into the allegations, Principal Backus was immediately removed as principal.

> *Philadelphia Inquirer,* June 25, 2006 (Patrick & Eichel, 2006).
> Edison Schools fires Jayne Gibbs, principal at Parry Middle School in Chester, Pennsylvania for allegedly changing student test answers in 2005. Eighth graders at the school said the principal had given them the answers to questions on the Pennsylvania System of School Assessment. Gibbs is also accused of exempting special-education students from testing, violating state and federal rules. Edison Schools also asks the state and district to investigate exemplary test results at Showalter Middle School, where Gibbs served as principal from 2003-04.

A third way in which educators manipulate test scores is by manipulating the test administration. These methods involve the most blatant forms of cheating educators can use to increase test scores, such as by changing student answers, giving students hints to test questions, or pointing out incorrect answers on the test. These manipulations also include any changes educators make to the test administration instructions, such as providing students with extra time to complete the test, rephrasing test items for students, giving students inappropriate accommodations, allowing students to work together on the test, or allowing students to use forbidden reference materials such as calculators or dictionaries on the test. Each of these methods serves to increase test scores without a corresponding increase in test scores, so each method is a manipulation.

The following three news report summaries provide examples of recent incidents in which educators have been caught allegedly manipulating the test administration:

> *Herald Tribune,* August 10, 2007 (Morris, 2007).
> Mary Cropsey, a third-grade teacher at Mills Elementary School in Manatee, Florida, is accused of tampering with student answer sheets on the Florida Comprehensive Assessment Test (FCAT). One student reports that Cropsey helped students on the test; another student reported hearing that the teacher gave students extra time to complete the exam. An investigation began after yet another student reported that she had not finished the exam, but the next day all the bubbles had been filled-in. If the allegations are proven true, Cropsey could lose her teaching certificate and even be charged with a crime.

> *Newsday,* June 24, 2007 (Hildebrand, 2007ab; Marcus, 2007ab)
> The entire Uniondale school district is placed on academic probation due to evidence of tampering with Regents Math A and B high school exams and the State Mathematics Assessments for grades 3-8 in 2005 and 2006. Faculty members reportedly allowed students to use calculators, which were not allowed on the exam (Marcus, 2007a). The New York Department of Education reports

> that complaints of test fraud have more than doubled over the past five years, with the department receiving 37 complaints in 2006. One dozen teachers and administrators accused of test fraud have faced hearings in front of the New York Professional Standards and Practices Board. Of those twelve cases, six cases resulted in revocation of professional certifications, two cases were cleared, and the remaining four cases remain under investigation. The number of complaints verified by the state has remained relatively steady, with between 9-16 in each of the past five years (Hildebrand, 2007a). An analysis of Uniondale's test scores found that 333 answers on the Regents Math A exam were altered, and 97% of the time they were changed to the correct answer. On the Regents Math B exam, 198 answers were changed, with 97% again being changed to the correct answer. On the 2005 8th grade math assessment, Uniondale students scored below average on 11 of the 14 easiest questions, but higher than average on 12 of the 13 most difficult items (Hildebrand, 2007b; Marcus, 2007b).
>
> *San Francisco Chronicle* May 13, 2007 (Asimov & Wallack, 2007) Teachers in at least 123 public schools have reportedly cheated for students on California's high-stakes tests between 2004-2006. In two-thirds of these cases, the schools admit that they had cheated. The cheating behaviors included (a) allowing students to use reference materials such as maps and flow charts during the test, (b) allowing students to use calculators, (c) helping students answer questions, and (d) erasing and changing student answers. California currently identifies potential misconduct by scanning answer sheets for suspicious erasures. Cheating is virtually ignored in schools in which cheating impacts less than 5% of tests are given. Schools in which cheating impacts more than 5% of the tests are not ranked and receive a note stating "adult irregularity in testing procedure" occurred.

The final way in which educators manipulate test scores is by manipulating score reports or performance standards. The most blatant manipulations in this category involve educators changing or removing student scores from official score records. Less obvious manipulations include educators changing demographic data so that scores from higher-ability students are added to specific lower-scoring subgroups or changing student identification numbers so that score trends cannot be calculated for lower-scoring students. A more nebulous manipulation in this category occurs when educators misrepresent test scores in score reports to make the scores appear better than they actually are. This includes changes made to lower the difficulty of the test or lower the cut-score for proficiency in order to make it appear as though student achievement has

increased. Because these methods cause test scores to increase without an increase in underlying student achievement, each of these methods is a manipulation.

The following three news report summaries provide examples of recent incidents of educators manipulating score reports or performance standards:

> *San Francisco Gate,* June 30, 2006 (Sturrock, 2006).
> According to researchers with Policy Analysis for California Education (PACE), California and eleven other states have inflated test outcomes by lowering the achievement standard students need to meet to be proficient in reading and math under the federal No Child Left Behind Act. The study describes large differences in results from state and national tests and outlines several reasons for these large differences. One of the reasons is that states sometimes lower their standards for what they deem proficient.
>
> *MSNBC,* April 17, 2006 (MSNBC, 2006)
> With permission from the federal government, nearly two million students' test scores are not counted when schools report progress by subgroups under the No Child Left Behind requirements. This is due to states being able to define the minimum number of students needed in a subgroup before scores are reported. In the past two years, almost half of all states have successfully petitioned the U.S. Department of Education to increase these minimums. An investigation concludes that about 1 out of every 14 test scores are not being counted under appropriate racial categories. The scores from more than 24,000 students in Missouri, 257,000 in Texas, and 400,000 in California are not being counted.
>
> *San Antonio Express News,* September 17, 1998 (Stinson, 1998)
> The Austin School District manipulated test results last spring to make it appear as if several schools performed better than they did, the Texas Education Agency (TEA) says. Commissioner of Education Mike Moses explained the trickery in this August 14 letter to the school district, stating, "… student identification number changes were submitted for students tested at (the schools), which resulted in the exclusion of those students from the accountability subset of TAAS results used to determine the 1998 accountability ratings." Administrators knew that by changing student identification numbers, the TEA would eliminate those students' scores from ratings calculations (Nichols & Berliner, 2004, p. 27).

Prevalence of Test Score Manipulations

In addition to providing examples of manipulations used by educators, the published news reports also provide a crude estimate of the prevalence of each

manipulation method. Figure 2.1 shows the number of published news reports about test manipulations each year from 1995 until 2007. Figure 2.2 shows the number of reports of manipulations by each state over that time period. These figures show that reports of manipulations are widespread and generally increasing over time, especially since the introduction of NCLB in 2002. Any conclusions made from these figures should be made cautiously, however. On one hand, published news reports can only be expected to represent the most interesting and, therefore, blatant incidents of test score manipulations. Also, news reports can only describe incidents in which educators were caught manipulating test scores. Therefore, these news reports may grossly underestimate the prevalence of manipulations. On the other hand, many of these news reports describe *allegations* of manipulations. Since follow-up reports are rarely written about these incidents, it is unknown how many of the reported allegations were eventually proven to be untrue. Therefore, these news reports may actually overestimate the prevalence of manipulations. Since it is unknown whether news reports over- or underestimate the actual prevalence of test score manipulations made by American public school educators, these reports can only provide evidence that educators have been reported to manipulate test scores and these reports are becoming more prevalent.

Recognizing the limitations in using published news reports, researchers have developed other methods to estimate the prevalence of test score manipulations. These methods include administering surveys to teachers and state officials; directly observing test administration procedures in classrooms; statistical detection; and other targeted research methods. Because each method has advantages and disadvantages, no one method provides the single best estimate of the prevalence of test score manipulations in American schools. Therefore, in order to best estimate the prevalence of manipulations, Table 2.2 synthesizes the estimates from research using each method. The estimated prevalence of each manipulation method is displayed using the taxonomy previously developed.

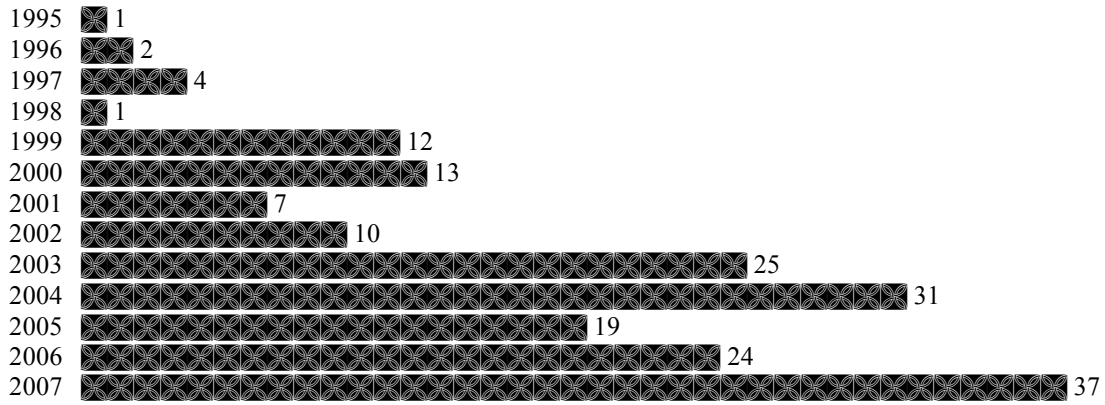


Figure 2.1 The number of published news reports on manipulations from 1995 – 2007.

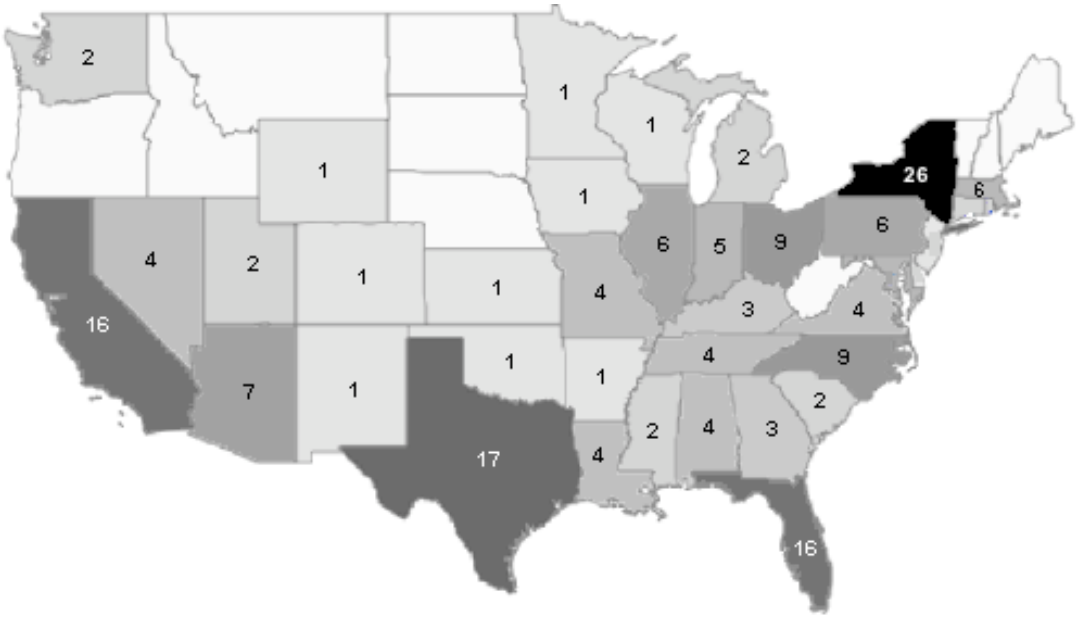


Figure 2.2 The number of published news reports from each state (1995 – 2007).

Table 2.2 Estimated Prevalence of Manipulations

	Teacher Surveys						State Survey	Stat Detect
	Gay (1990)	Shepard & Dougherty (1991)	KD-LG[1] (1992)	Nolen et al.[2] (1992)	Pedulla et al. (2003)	Lai & Waltman (2007)	Mehrens et al.[4] (1993)	JL[6] (2003)
Sample size	N=168	N=360	N=74	N=1881	N=4195	N=1338	N=46	
Sample location	NC	2 districts	LA	AZ	47 states	IA	States	Chicago
Manipulate Teaching Philosophy or Process								
Making copies of test prior to test administration	---	---	---	---	---	---	10%	---
Practicing with items identical or similar to the test	---	11%	8%	10%; 9%	32%	11%	33%	---
Practice with last year's (alternate form) test items	---	---	66%	12%; 9%	---	15%	---	---
Practice with items of the same format as the test	---	---	---	---	---	---	---	---
Use commercial test preparation packages	---	---	53%	41%; 12%	---	---	---	---
Teaching test-taking skills; test-wiseness	---	---	76%	60%; 36%	76%	82%	---	---
Teaching content from specific test items	23%	41%	---	23%	6%	---	---	---
Focusing resources on students closest to proficiency	---	---	---	---	13%	---	---	---
Primarily teaching content found on the test	---	---	39%	66%; 22%	69%	18-22%	---	---
Manipulate Examinee Pool								
Excluding students from testing	---	8-13%	0%	---	---	---	---	---
Bribing or paying students to increase test scores	---	---	---	---	14%[3]	---	---	---
Having high-scoring students take test multiple times	---	---	---	---	---	---	---	---
Providing inappropriate special education placement	---	---	---	---	---	---	---	---
Increasing caloric content of school meals	---	---	---	---	---	---	---	---
Manipulate Score Reports or Standards								
Changing student test scores on official records	---	---	---	---	---	---	---	---
Providing false IDs so scores won't count	---	---	---	---	---	---	---	---
Misrepresenting data	---	---	---	---	---	---	---	---
Changing criteria for proficiency; making test easier	---	---	---	---	---	---	---	---
Manipulate Test Administration								
Changing student test scores on official records	2%	6%	1.4%	---	2%	---	---	4-5%

Table 2.2 Continued

	Gay (1990)	Shepard & Dougherty (1991)	KD-LG[1] (1992)	Nolen et al.[2] (1992)	Pedulla et al. (2003)	Lai & Waltman (2007)	Mehrens et al.[4] (1993)	JL[6] (2003)
Manipulate Test Administration								
Sanitizing answer sheets (cleaning before scoring)	---	---	88%[5]	---	---	---	---	---
Giving students answers	---	8%	---	---	---	---	---	---
Checking or pointing out incorrect answers	10%	---	---	---	11%	---	---	---
Giving students (non)verbal hints on test items	14%	23%	0%	---	11%	---	---	---
Not following test administration procedures exactly	4%	---	---	8%; 6%	---	---	---	---
Rephrasing test items for students	---	18%	10%	---	---	---	---	---
Allowing students to work together during testing	2%	---	---	---	---	---	---	---
Ignoring students who are cheating	---	---	---	---	---	---	---	---
Giving students additional examples	---	---	---	28%; 16%	---	---	---	---
Providing students extra time	14%	20%	1%	8%; 3%	15%	---	---	---
Reading items that are to be read by students	---	14%	---	---	---	---	---	---
Answering questions about test content	---	12%	---	---	---	---	---	---
Providing students with reference materials or tools	1%	---	---	---	---	---	---	---
Having students fill-in unanswered items	---	---	---	---	---	---	---	---
Providing inappropriate accommodations to students	---	---	---	---	---	---	---	---
Review skills that will be on tomorrow's test	---	---	---	44%; 30%	---	---	---	---

Notes: 1 Kher-Durlabhji & Lacina-Gifford (1992) sampled *preservice* teachers about what manipulations they planned on using

2 Numbers represent separate responses from elementary; secondary school teachers

3 Actual survey item was worded "give prizes to reward students"

4 Numbers represent the percent of states reporting specific incidents during the 1989-1990 academic year

5 Actual survey item was worded "check student's completed answer sheets"

6 Jacob and Levitt (2003) attempted to detect the percentage of teachers who manipulate answer sheets for their students

Prevalence Estimates Based on Teacher Surveys

The most frequently used method to estimate the prevalence of test score manipulations involves administering a survey to teachers or school administrators. These surveys, which are usually byproducts of larger research into test preparation activities or the impact of high-stakes testing on instruction, typically ask teachers to indicate which manipulation methods they use to increase test scores.

Due to concerns that teachers will be reluctant in admitting to using several of the more blatant forms of manipulation, many surveys also ask teachers to report if they are aware of other teachers in their schools manipulating test scores. Table 2.2 displays the results of six of these teacher surveys. The numbers in the table represent the percentage of teachers responding to each survey who report that either they or other teachers in their schools use each method of manipulation.

The first teacher surveys that asked about the use of a wide range of manipulations were administered in the early 1990s. Gay (1990) administered a survey to 168 North Carolina teachers in grades 3 through 8. Gay found 35% of respondents reported participating in or being aware of testing irregularities in their schools (p. 4). These testing irregularities, defined by eight specific examples, all fit within the definition and taxonomy of test score manipulations. According to the results, 23% of teachers reported manipulating the teaching process by copying the test and teaching its contents to students prior to test administration. Fewer respondents reported manipulating the test administration, with 15% adding extra time to the test publisher's time limits, 14% coaching students on the test by giving verbal or nonverbal hints, 10% calling attention to incorrect student answers, 4% changing the publisher's test administration directions, and 2% leaving students unsupervised during testing. The most blatant form of manipulation, changing student answers, was reported by 2% of the respondents. Gay also noted that one respondent reported an incident in which a teacher

encouraged students to use reference materials during a writing test and another respondent admitted to checking student responses to "be certain that her students answered as they had been taught" (p. 4). In reporting the survey results, Gay suggested that the estimated prevalence of each test score manipulation "may only be the tip of the iceberg" (p. 3) and reported that 43% of respondents reported a belief that test manipulations were increasing among teachers.

The next year, Shepard and Dougherty (1991) administered a similar survey to 360 teachers from two American school districts as part of a larger study to determine the effect of the high-stakes tests on instruction and student learning. In this survey, teachers were asked to report the frequency with which "controversial testing practices" happened in their schools. These testing practices all fit within the definition of test score manipulations. Supporting the results from Gay's survey, the researchers found that manipulating the teaching process was the most prevalent form of manipulation, with 41% of respondents occasionally or frequently giving highly similar items to students for practice and 11% practicing with items from the actual test. Fewer teachers reported manipulating the test administration, with 23% providing hints to students during testing, 20% giving students more time than the test directions call for, 18% rephrasing test questions for students, 14% reading test questions that were supposed to be read by students, 12% answering questions about test content during test administration, and 8% giving answers to students. Shepard and Dougherty also found that 6% of respondents reported the most blatant manipulation of changing incorrect answers to correct ones on student answer sheets. Shepard and Dougherty also found evidence of educators manipulating the examinee pool. 13% of respondents did not administer the test to students who would have trouble and 8% encouraged lower-ability students to be absent on the days of the test.

The following year, Hall and Kleine (1992) surveyed 220 Oklahoma public school teachers and found that 55% reported awareness of fellow teachers cheating on

tests for their students. While the term *cheating* was not defined on the survey, the results would certainly provide another estimate of the prevalence of activities designed to increase test scores without a corresponding increase in student achievement.

Rather than asking teachers to report manipulations they currently use, Kher-Durlabhji and Lacina-Gifford (1992) surveyed 74 pre-service teachers from Louisiana to determine they types of test preparation and administration activities they plan to use in teaching. Once again, the most prevalent forms of manipulation were methods used to manipulate the teaching process or philosophy. 76% of these pre-service teachers intended to spend instructional time on teaching test-taking skills, 66% intended to practice with previous years' test questions, 53% intended to use commercial test preparation packages, 39% intended to teach only content found on the test, and 8% intended to practice with items from the actual test. Also supporting previous survey results, fewer pre-service teachers intended to manipulate the test administration. 10% indicated that they plan to rephrase test items for their students and 1% planned to provide students extra time to complete the test. Not a single pre-service teacher intended to give students hints on test items. Also, no respondents intended to manipulate the examinee pool in order to increase test scores. It should be noted that while 88% of pre-service teachers intended to "check students' completed answer sheets," this does not necessarily imply they will take any action after checking the answer sheets. Thus, this does not describe a manipulation of test scores.

While the previous surveys were administered to small samples of educators, Nolen, Haladyna, and Haas (1992) administered a series of large-scale surveys to Arizona elementary and secondary teachers to determine their uses of test scores. As part of this research, 1,881 Arizona teachers were asked to indicate the test preparation practices they always or usually employ and which test administration activities were common or very common in their classrooms. In addition to finding that elementary school teachers were more likely to report manipulations than secondary teachers, the researchers once again

found that manipulations of teaching philosophy or process were most frequently reported. The prevalence of specific manipulations in this category ranged from 66% of elementary school teachers teaching only content found on the test to 9% of secondary teachers giving their students items from the actual test as practice. As further evidence of the prevalence of manipulations of the teaching process, 23% of elementary teachers responding to the survey indicated a belief that administrators required them to spend class time on test preparation activities and 33% reported spending more time on test preparation than was required. Also, 7% of elementary teachers, 10% of secondary teachers, and 3% of school administrators surveyed reported that teachers are encouraged to raise test scores by teaching items from the actual test. Manipulations of the test administration were less prevalent, with 44% of elementary school teachers admitting to reviewing the tested content and skills immediately before testing and 3% of secondary teachers providing extra time for students to complete the test.

Lai and Waltman (2007) administered a similar survey to a sample of 1,338 teachers from 125 public schools in Iowa. The estimated prevalence of manipulations of the teaching process or philosophy ranged from a median of 82% of teachers within a school teaching test-taking skills to their students to a median of 11% of teachers within a school giving actual items from the test to their students as practice. The researchers found that as many as 50% of teachers within a school admitted to using actual items from the test as practice and as many as 67% of teachers within a school allowed students to practice with the alternate form of the test. More troubling was the finding that "unexpectedly high percentages of teachers rated practicing with exactly the same test that will be administered this year as being 'very ethical'" (p. 12).

While the surveys from Nolen, Haladyna, and Haas (1992) and Lai and Waltman (2007) were administered to large samples of teachers, the fact that their surveys were administered to teachers in only one state may limit the generalizations that can be made. To address this problem, the National Board on Educational Testing and Public Policy

funded a survey of 4,195 teachers from every state except Iowa, Oregon, and Idaho (Pedulla et al., 2003). The survey, designed to measure teacher attitudes towards state testing programs, asked teachers how their state's testing program impacted their classroom instruction and test preparation activities. Results from this survey once again showed that manipulations of the teaching process or philosophy were the most prevalent form of manipulation used by teachers. More than two-thirds of the respondents reported teaching test-taking strategies to their students and teaching only content found on the actual test, and one third reported practicing with items identical or similar to those found on the actual test. Less prevalent ways of manipulating the teaching process or philosophy included 13% of teacher focusing instructional resources on students who were closest to achieving a proficient score and 6% teaching content from specific test items. To manipulate the test administration, 15% give students extra time to complete the test, 11% give hints to students, and 11% point out incorrect answers to students. Only 2% admitted to the most blatant manipulation of changing answers on student answer sheets. In an effort to manipulate the examinee pool, 14% of teachers reported giving prizes to reward students for higher test scores. As a general estimate of the prevalence of test score manipulations, 38% of the respondents indicated that teachers in their schools have found ways to raise state-mandated test scores without really improving learning.

While the results from this method of administering surveys to teachers seem to converge to provide estimates of the prevalence of each manipulation method, the results should be interpreted carefully. As stated earlier, results from the surveys of Gay (1990), Shepard and Dougherty (1991), Hall and Kleine (1992), and Kher-Durlabhji and Lacina-Gifford (1992) are based on extremely limited sample sizes that may not generalize beyond the sample. While Nolen, Haladyna, and Haas (1992) and Lai and Waltman (2007) administered their surveys to much larger samples of teachers, the fact that each of these surveys was administered to teachers within a single state may also limit any

generalizations. This concern is further supported by the low 42% response rate from they survey administered by Nolen, Haladyna, and Haas (1992, p. 10). Results from the survey from the National Board on Educational Testing and Public Policy (Pedulla et al., 2003), while collected from a much more representative sample of teachers, should also be cautiously interpreted due to the reported 35% response rate. Further encouraging a cautious interpretation, the demographics of the sample did not match national demographics, with the sample containing more males, more experienced teachers, and more English teachers than the national average (Pedulla et al., 2003, p. 138).

Results should also be interpreted carefully due to the impact of social desirability bias on surveys. Social desirability bias is the inclination one has to respond to survey items in a manner that will be viewed favorably by others (Paulhus, 1991). In a study of student cheating on achievement tests, Scheers and Dayton (1987) found this effect, concluding that survey responses underestimated the actual extent of cheating behaviors. In a similar study of student cheating, Nelson and Schafer (1986) found the exact opposite effect. These researchers found that surveyed responses overestimated actual cheating incidents. While Eve and Bromley (1981) developed *culture-conflict theory* to explain these opposite results, the fact that response bias cannot be predicted limits the accuracy of manipulation prevalence estimates from these surveys.

Finally, with the exception of Lai and Waltman (2007), each survey treated individual teachers as independent units of analysis. Since administrative decisions might be made to influence teachers' decisions to manipulate test scores, teachers are not independent units (Lai & Waltman, 2007, p. 14) perhaps a more accurate estimate could be made using schools, school districts, or states as the units of analysis.

Prevalence Estimates Based on State Surveys

Rather than surveying individual educators to estimate the prevalence of manipulations, Mehrens, Phillips, and Schram (1993) surveyed 46 state departments of

education on behalf of the National Council on Measurement in Education. The researchers asked state officials to indicate the number of test security incident reports they received involving their state tests. Of the 41 states with testing programs at the time, 36 (88%) indicated receiving reports of test security breaches, with 28 states (68%) receiving such reports during the 1989-1990 academic year. Although some states completed the survey incorrectly (p. 5), the researchers found that the 41 states received an average of 6.34 reports of security breaches in that single academic year.

Supporting the results from the teacher surveys, Mehrens, Phillips, and Schram (1993) found that manipulations of the teaching process or philosophy were the most prevalent form of test score manipulation. Four states (10%) received at least one report regarding missing test materials in 1989-1990. The states reported that the test materials went missing due to unauthorized personnel having access to the materials or due to shipping irregularities. Ten states (24%) received reports of teachers using items very similar to the actual test items as practice and 14 states (34%) received reports of teachers providing their students with test questions in advance.

States received fewer reports of manipulations of the test administration during the 1989-1990 academic year. While 18 states (44%) received reports of teachers failing to follow the written test administration procedures, only 7 states (17%) received reports of educators erasing student answers. Only two states (5%) received reports of the most blatant form of manipulation – teachers marking answers on student answer sheets. As further evidence of the prevalence of test score manipulations, 10 states (24%) received reports of "dramatic increases" in average test scores for schools or school districts (p. 7).

In another survey of state departments of education, the test security firm Caveon (Sorensen, 2006) asked states to indicate the number of times in the past two years that they have taken formal action because cheating was either confirmed or suspected. While 6 of the 34 states surveyed (18%) reported no formal actions taking place, 13 states (38%) took formal action up to 3 times, 5 states (15%) took formal action 4 to 6 times, 2

states (6%) of states took formal action 7 to 9 times, and 8 states (24%) took formal action more than 10 times in two years.

As a more indirect estimate of manipulation prevalence, state officials were also asked to rate the importance of various test security threats to their state testing programs. Yet again manipulations of the test administration were perceived to represent a greater threat to state testing programs than manipulations to the teaching process or philosophy. States rated lost test materials, inappropriate administrative pressure put on teachers to increase test scores, and teacher coaching of students based on prior knowledge of test questions as the most important perceived threats. Manipulations of the test administration, including students working together on the test, teachers providing answers to students during testing, and educators changing student answers after testing were perceived to be less important.

While the results from these state surveys do support the results from teacher surveys in showing that manipulations do happen and manipulations of the teaching process or philosophy are more prevalent than manipulations of the test administration, the reported prevalence estimates must once again be interpreted cautiously. These surveys only indicate the number of manipulation incidents reported to state officials. In his survey of 60 teachers, Gay (1990) found that only 20% were willing to report testing irregularities to their school administrators (p. 4). If we assume that school administrators are similarly reluctant to report testing irregularities to district and state officials, then the results from these state surveys could greatly underestimate the actual prevalence of test score manipulations.

Prevalence Estimates Based on Direct Observation

To overcome the potential impact of response bias in teacher surveys and the limitations of state surveys, some researchers have directly observed schools and classrooms to determine the prevalence of test score manipulations. Due to its resource

intensiveness, observational studies of classroom test administration have been rare. In 1981, Horne and Gary observed the administration of a standardized test in 16 elementary school classrooms. The researchers found more than half of the teachers varied from the written test directions with 6 (38%) of the teachers "purposely and consciously [manipulating] the test administration procedures" (Horne & Gary, 1981, p. 12).

White, Taylor, Carcelli, and Eldred (1981) similarly observed the administration of a standardized test in 38 Utah classrooms. Supporting the previous study, the researcher found nearly half of the teachers did not follow the test administration directions exactly. The study found 46% of teachers failed to follow the exact wording in presenting test questions as stated in the test manual, 41% changed the wording of the test directions to a "vocabulary more familiar to students," and only 50% "refrained from repeating a test question unless the directions specified to do so" (White et al., 1981).

Wodtke, Harper, Schommer, and Brunelli (1989) continued this method of research by observing the test administration practices of ten kindergarten classrooms. The researcher found, once again, that teachers manipulated the test administration. The researchers observed teachers failing to follow the time limit directions found in the test manual in 27% of the testing sessions. The researchers also recorded 21 unauthorized item repetitions, 40 incidents of teachers "cueing correct answers," and 149 "significant procedural variations," including rephrasing test questions and failing to disseminate practice test booklets (p. 228). The researchers concluded that administrations of standardized tests are manipulated so much as to render them incomparable (Wodtke et al., 1989).

While the method of direct observation has advantages over teacher and state surveys, the method is not without faults. First, due to its resource-intensiveness, only a small (and usually nonrepresentative) sample of classrooms can be observed. Second, direct observation methods may be subject to both the *Hawthorne Effect* and the *observer-expectancy effect*. The Hawthorne Effect suggests that subjects in direct

observation studies temporarily change their behavior due to their knowledge that they are part of a study (BJA, 2007). Thus, in studies of test administration manipulations, the Hawthorne Effect suggests that estimates based on direct observation may underestimate the actual prevalence. The observer-expectancy effect, on the other hand, is "a cognitive bias that occurs when a researcher expects a given result and therefore unconsciously manipulates an experiment or misinterprets data in order to find it" (Ferguson, 2007). If the researchers expected to observe test administration manipulations and if the observer-expectancy effect is real, then results from these direct observation studies might overestimate the actual prevalence.

Prevalence Estimates Based on Statistical Detection

In an effort to eliminate the response bias in surveys and potential biases in direct observation, statistical detection methods have also been used to estimate the prevalence of manipulations. These methods, detailed in Appendix B and Cizek (1999), use statistical analyses of student answer sheets in order to detect potential manipulations. First developed with the intention of detecting students who cheat on tests, these methods attempt to find examinee answer sheets with unusually large score gains or wild score fluctuations (Perlman, 1985); statistically improbable numbers of erasures (Qualls, 2001); or unusual patterns of answers (Advanced Psychometrics, 1993; Angoff, 1974; Anikeeff, 1954; Bay, 1995; Bellezza & Bellezza, 1989; Bird, 1927, 1929; Cizek, 1999; Drasgow, Levine, & Williams, 1985; Frary, 1977; Hanson & Brennan, 1987; Jacob & Levitt, 2004; Karabatsos, 2003; Kvam, 1996; Levine & Drasgow, 1979, 1988; Meijer & Sijtsma, 2001; Roberts, 1987; Saupe, 1960; Sotaridona & Meijer, 2001, 2002; van der Linden & Sotaridona, 2002; Wesolowsky, 2000; Wollack, 1997).

In 1985, Perlman analyzed student answer sheets from a standardized test administration in Chicago Public Schools. Discovering some schools had unusually large score gains and unusually high numbers of answer sheet erasures, Perlman hypothesized

that educators at these "suspect" schools may have manipulated the test administration process. In an effort to test his hypothesis, Perlman retested 23 of these suspect schools and 17 control schools that were not suspected of manipulating scores. The results of the retesting led him to state "clearly the suspect schools did much worse on the retest than the [control] schools," and conclude that, "it's possible that we may have underestimated the extent of cheating [manipulations of the test administration] at some schools" (Perlman, 1985, pp. 4-5).

Almost two decades later, Jacob and Levitt (2003, 2004) conducted another analysis of answer sheets from Chicago Public Schools. The researchers developed a statistical index to detect educators who manipulate student responses on tests by supplying them with answers or erasing and changing student responses. Applying their index to analyze 8 years of answer sheets from the Chicago Public Schools' administration of the *Iowa Tests of Basic Skills* (ITBS), the researchers concluded, "Empirically, we detect cheating [manipulation of answer sheets] in approximately 4 to 5 percent of the classes in our sample" (2004, p. 846). They also found that between 1.1% and 2.1% of educators in their sample manipulated answer sheets on any particular ITBS subtest and 3.4% to 5.6% of educators manipulated answer sheets on at least one ITBS subtest. Further describing the prevalence, Jacob and Levitt (2003) conclude that educators found to manipulate answer sheets on one test were 10 times more likely to manipulate answer sheets on other tests and educators found to manipulate answer sheets one time were 9 times more likely to do so again in the future (p. 73).

In order to validate their estimate of 4% to 5% of educators manipulating the test administration by changing student answer sheets, the researchers retested 117 Chicago classrooms in 2002. The educators in these classrooms included "cheaters" whose classrooms experienced large score gains and showed evidence of unusual response patterns; "bad teachers who cheat" whose classrooms had unusual response patterns but did not experience large score gains; "anonymous tips" whose classrooms were not

identified by the statistical index but who were accused of cheating; "effective teachers" whose classrooms experienced large score gains with no evidence of unusual response patterns or manipulations; and "randomly selected" educators whose classrooms were not suspected of manipulations. Scores from the students in the "effective teachers" classrooms increased on the retest, while the "randomly selected" classrooms experienced a small decline of 2.3 standard score units. The "cheaters," "bad teachers who cheat," and "anonymous tips" classrooms experienced a large decline in score of 16.2, 8.8, and 6.8 standard score units, respectively. One of the classrooms taught by a "cheater" experienced a loss of 54 standard score units on the retest – a loss roughly equivalent to three full grade equivalent units on the ITBS. Based on the results of this retesting, Jacob and Levitt (2004) expressed confidence in their estimate of the prevalence of manipulations of student answer sheets.

Using a different statistical index, Wesolowsky (2000) analyzed answer sheets from the 2005-2006 administration of the *Texas Assessment of Knowledge and Skills* (TAKS). While not providing an overall estimate of the prevalence of manipulations of the test administration, Wesolowsky found that scores from more than 50,000 students showed evidence of irregularities that could include students copying answers from other students or educators doctoring student answer sheets. Additionally, the analysis found 112 schools in which at least 10% of student answer sheets were identified as potentially being manipulated. Expressing confidence in his index's conservative estimate of manipulations, Wesolowsky stated, "The evidence of substantial cheating is beyond any reasonable doubt" (Benton & Hacker, 2007a, 2007b).

While these statistical detection methods have advantages over the survey and direct observation methods, they do have limitations. First, statistical indices can only detect manipulations of answer sheets. They cannot detect other manipulations of the test administration or manipulations of the teaching process or philosophy, examinee pool, or score reports or standards. Second, these methods can only detect *possible* manipulations

of answer sheets. Some students, classrooms, and schools can experience legitimate large gains in test scores. Likewise, an unusual response string from an examinee or group of examinees does not necessarily mean that the answer sheets have been manipulated. For these reasons, statistical detection indices might overestimate the actual prevalence of test administration manipulations.

On the other hand, many of the statistical indices have been shown to be ineffective in detecting simulated manipulations of answer sheets (Chason & Maller, 1996; Iwamoto, Ningester, & Luecht, 1996). Demonstrating this ineffectiveness, test security firm Caveon analyzed a simulated data set using six of their indices to detect unusual response strings. The data set was simulated so that 3,283 answer sheets had been manipulated by teachers (teachers changing answers from incorrect to correct). The data were also simulated to include 5 "schools" of answer sheets had been manipulated by administrators (principals or school personnel changing answers for all students). With this data set, the firm's indices were only able to detect 41 (1.2%) of the simulated manipulated answer sheets and none of the five simulated schools (Impara, Kingsbury, Maynes, & Fitzgerald, 2005). Due to their poor ability in detecting manipulations, statistical indices might actually underestimate the actual prevalence of test administration manipulations.

Prevalence Estimates Based on Targeted Research

While the survey, direct observation, and statistical detection methods provide estimates of the prevalence of manipulations of the teaching process or philosophy and administration, they rarely provide information regarding the prevalence of manipulations of the examinee pool. To fill this gap, researchers have designed targeted research studies. These studies find that schools, school districts, and states do manipulate the examinee pool in order to increase test scores.

Educators who manipulate the examinee pool usually do so by excluding lower-ability students from testing. One way in which educators do this is by suspending or otherwise punishing lower-ability students during the test administration period so they are not able to participate. Figlio (2005) hypothesized that during the test administration period, schools with accountability systems would give low-ability students harsher penalties (longer suspensions) than they would give to higher-ability students. Figlio supported his hypothesis by citing evidence that students who receive suspensions of at least one week in length were twice as likely to miss the test administration and the make-up testing dates as students who receive shorter suspensions (p. 3). To test his hypothesis, Figlio analyzed test and discipline data from 41,803 students in Florida school districts during the four years following the introduction of the state's high-stakes *Florida Comprehensive Assessment Test* (FCAT). In his analysis, Figlio compared the lengths of suspensions given to at least two students for the same incident. By classifying students as low- or high-ability based on previous years' test scores, Figlio was able to compare the suspension lengths to see if ability had an influence. Figlio found that, "While schools always tend to assign harsher punishments to lower-ability students than to higher-performing students throughout the year, this gap grows substantially during the testing window. Moreover, this testing window-related gap is only observed for students in testing grades" (pp. 4-5). Figlio also found that given two students suspended for the same incident during the test administration period, low-ability students were 12.3% less likely to take the FCAT than higher-ability students (p. 19).

In stating his conclusions, Figlio did address potential concerns. The first concern is that maybe low-ability students are more likely to be suspended during the test administration period because they want to avoid testing. The second concern is that perhaps low-ability students are more likely to cause the incident or be worse offenders, so therefore they are more likely to receive longer suspensions. The researcher addresses these concerns by reporting that low-ability students tend to get suspended at similar

rates, relative to high-ability students, during the test administration period as in other times of the year.

Another way in which educators can manipulate the examinee pool is by inappropriately classifying low-ability students as disabled. Before NCLB, states such as Florida had rules in which disabled students were exempt from taking the state test (Figlio & Getzler, 2002). States could, then, improve their test scores by simply classifying their lowest-scoring students as disabled. To test the hypothesis that states do, in fact, manipulate the examinee pool in this way, Figlio and Getzler (2002) analyzed 9 years of test data from six school districts in Florida. The researchers found that in the five years before the introduction of the state high-stakes accountability system, between 7.3% and 8.8% of students were classified as disabled. In the three years following the introduction of the accountability system, the classification rate increased each year from 9.4% to 9.6% to 10.8%. Controlling for this nearly linear increase in classification rates, the researchers found that the introduction of the high-stakes accountability system led to a 5.6% increase in the likelihood that a student was classified as disabled. According to the researcher, "the introduction of FCAT testing is associated with a more than 50% higher rate of disability classification" (p. 9).

To address the concern that perhaps Florida is unique in its manipulations of the examinee pool, Cullen and Reback (2002) and Jacob (2007) conducted similar studies in Texas to determine if lower-ability students were inappropriately classified as disabled in order to increase test scores. Both studies concluded that lower-ability students were more likely to be exempt from testing and that educators do, in fact, manipulate the examinee pool through inappropriate disability classification.

Another way in which educators manipulate the examinee pool is by disproportionately focusing instructional resources on students who have the best chance to improve the school's overall test scores. If school performance is determined by the percent of students earning a proficient score on a test, then educators may be tempted to

focus their attention on students who earned scores just below proficient the previous year at the expense of students who scored extremely high or low on the previous year's test. Neal and Schanzenbach (2007) tested this hypothesis in a study of test data from Chicago Public Schools. The data came from 1998, following the introduction of a high-stakes accountability system. This accountability system evaluated schools by examining the percentage of students in each school who earned a proficient score. The researchers found that students near the middle of the achievement distribution achieved at a higher level after the accountability system was introduced. The also found that students at the low-end of the achievement distribution achieved at the same or even a lower level after the accountability system was introduced. The researchers found that "for at least the bottom 20% of students, there is little evidence of significant gains and a possibility of lower than expected scores" (p. 27) following the introduction of the accountability system. This seems to support the hypothesis that educators manipulate the examinee pool in Chicago.

Reback (2007) conducted a similar analysis on test data from Texas. The researcher hypothesized that a state accountability system "increases incentives for schools to improve the performance of students who are on the margin of passing but does not increase short-run incentives for schools to improve other students' performance" (p. 1). While Reback found that accountability systems do improve overall student achievement, most of these gains were realized for students whose achievement levels were closest to the cut-scores. The researchers found that "other students only make greater than expected gains in this situation if their own performance is particularly important for their schools' rating" (p. 33). These conclusions support the belief that educators in Texas manipulate the examinee pool in order to increase test scores.

An unusual way in which educators have been shown to manipulate the examinee pool is through, surprisingly enough, the school lunch program. Figlio and Winicki (2003) designed a study to target this specific manipulation method. After claiming "the

link between nutrition and cognitive ability has been well established" (p. 382), the researchers examined the school lunch menus from Virginia public schools during the 1999-2000 academic year. From this data, the researchers were able to conclude, "schools threatened with accountability sanctions increase the caloric content of their lunches on testing days in an apparent attempt to boost short-term student cognitive performance" (p. 381). Moreover, the researchers found evidence that this manipulation of the school lunch program was effective in raising test scores. The researchers found that schools that increased the caloric content of their lunches on testing days by 100 experienced a 7% increase in the pass rate of students on the mathematics test (p. 392).

Another targeted research method used to determine the prevalence of manipulations of score reports or score standards is through analyses of state testing programs and their cut-scores for proficiency. By lowering their proficiency standards or making their state tests easier, states can inflate test scores (or test score comparisons) without actually increasing student achievement. Some studies that attempt to document these manipulations simply compare the tests and proficiency standards from each state. In a report for *CBS News*, Wallace (2007) concluded that the large variability in state proficiency rates was due, primarily, to differences in the difficulty of state tests and the cut scores used for proficiency. Sturrock (2006), reporting for the *San Francisco Gate*, suggested that differences in results from California's state test and the National Assessment of Educational Progress (NAEP) provided evidence that California had manipulated scoring standards to make it appear as though student achievement had increased. Although the methods and conclusions from these analyses are oftentimes questionable (as will be discussed later), they can provide evidence of educators manipulating test scores

Why Do Educators Manipulate Test Scores?

The previously discussed surveys, observational studies, statistical detection studies, and other targeted research results provide evidence that educators are manipulating test scores and that these manipulations can have a significant impact on test scores. In order to prevent this behavior, it must first be understood why educators manipulate test scores. Research into cheating and inappropriate test preparation activities suggests at least four reasons why educators would manipulate test scores: (1) educators are former students, (2) pressure from high-stakes test accountability systems, (3) a lack of understanding of what behaviors are inappropriate, and (4) a lack of oversight and policies to deter manipulations. While these reasons are neither mutually exclusive nor exhaustive, they can provide insight into how manipulations could possibly be deterred.

Educators Are Former Students

Over the last century, researchers have published more than one hundred studies on the prevalence of student cheating on exams (Cizek, 1999, 2003). Estimates have ranged from 5% of examinees cheating on any particular occasion (Bellezza & Bellezza, 1989) to 75% of students admitting to some form of cheating before graduating high school (Impara, Kingsbury, Maynes, & Fitzgerald, 2005) to more than 80% of American undergraduate students admitting to cheating during college (Passow, Mayhew, Finelli, Harding, & Carpenter, 2006). A 2006 survey of 36,000 students by the Josephson Institute (2006) found that 60% of examinees cheated on a test during the past year; 35% cheated two or more times; 33% admitted using the internet to plagiarize an assignment; and 27% admitted to lying on at least one question on this survey.

Educators are former students. If students cheat on tests and students become educators, then educators may continue that cheating behavior to manipulate test scores in their classrooms. In a study of 5,280 students across nine academic majors, Bowers

(1964) found that 52% of undergraduate education majors reported cheating on a test during college. Cizek (1999) described an observational study conducted in the 1920s in which 110 women about to begin student teaching were allowed to score their own tests in a college-level education course. The researchers found that 30 (27%) of the women cheated by changing their answers during the self-scoring, with 4 (3.6%) of the women changing more than 10 answers. Based on these studies, it is safe to assume that many educators have cheated on a test at least once in their academic careers. Because research has found that the decision to cheat in college is correlated with the decision to later engage in other unethical behaviors in the workplace (Crown and Spiller, 1998), educators who cheated as students may choose to manipulate test scores in their classrooms.

In replicating a 30-year old large-scale study of undergraduate student cheating, McCabe, Trevino, and Butterfield (2001) found that while the overall prevalence of cheating increased only modestly, the prevalence of the most blatant forms of test cheating "increased significantly" (p. 221). In explaining possible causes of cheating, the researchers found that contextual factors (peer cheating, peer disapproval of cheating, or perceived severity of penalties for cheating) were significantly more influential than individual factors (age, gender, academic ability, or participation in extracurricular activities. The researchers explained:

> Students who might otherwise complete their work honestly observe cheating by others and convince themselves they cannot afford to be disadvantaged by students who cheat and go unreported or unpunished. Although many find it distasteful, they too begin cheating to level the playing field. The strong influence of peers' behavior may suggest that academic dishonesty not only is learned from observing the behavior of peers, but that peers' behavior provides a kind of normative support for cheating. The fact that others are cheating may also suggest that, in such a climate, the non-cheater feels left at a disadvantage. Thus cheating may come to be viewed as an acceptable way of getting and staying ahead (pp. 220-222).

This reasoning is supported by a 1997 report from *Who's Who Among American High School Students* on academic cheating (Newberger, 1997). According to the report, 92% of confessed cheaters declared that they had never been caught. Newberger suggested that students cheat not only to keep up with other cheating students, but also because they think they can get away with it. Crown and Spiller (1998) also found that while lower-ability students were slightly more likely to cheat, contextual factors such as a school's lack of an honor code or weak penalties for cheating increase the likelihood of cheating. The researchers also concluded, "The amount of unflattering attention the popular press gives to the students reporting high percentage levels of collegiate cheating could lead many students to the conclusion that they must cheat just to keep up with their peers" (pp. 684-695).

Cizek (2003) suggests that these contextual factors also influence an *educator's* decision to manipulate test scores. As he describes, "Because so much of that cheating went undetected and unpunished, and because they can easily put themselves in the position of examinees desperate to pass a test, those who give tests may often be tempted to turn a blind eye to cheating (pp. 6-7). Jacob and Levitt (2003) found evidence of these contextual influences in their study of educators who change student answers on tests. In addition to finding that younger educators were more likely to cheat than older educators, the researchers found that educators in classrooms that performed poorly on the previous year's exam, and educators in classrooms with higher poverty rates and more minority students were more likely to cheat. These contextual factors along with many educators' histories of cheating as students may explain why many educators manipulate test scores.

Pressure From State Accountability Systems

Other researchers suggest that high-stakes accountability systems are the reason why educators manipulate test scores. Some studies have found that the pressure felt by educators, whether real or perceived, to improve test scores causes them to manipulate

test scores. Hatch and Freeman (1988) found this when they interviewed kindergarten teachers in Ohio. 67% of the interviewed teachers reported "implementing instructional practices in their classrooms that they considered to be antithetical to the learning needs of young children; they did this because of the demands of parents and the district and state accountability systems" (Hatch & Freeman, 1988, p. 146). Hamilton and Stecher (2006) also found this in their survey of 2,628 math teachers and 262 principals from elementary and middle schools in California, Georgia, and Pennsylvania. The researchers found that 79-92% of teachers felt a great deal of pressure to improve scores on the state mathematics test. Because of this pressure, 19% - 78% of teachers manipulated the teaching philosophy or process by: (a) focusing more on topics emphasized on the state test, (b) emphasizing the formats and styles of test items in instruction, (c) spending more time teaching general test-taking strategies, (d) focusing more effort on students who are close to proficient, or (e) offering more assistance outside of school to help students who are not proficient (p. 22). More than half of the principals also responded to this pressure by encouraging teachers to manipulate the teaching philosophy or process by: (a) distributing commercial test preparation materials, (b) encouraging or requiring teachers to spend more time on tested subjects and less time on other subjects, or (c) encouraging teachers to focus their efforts on students close to meeting the standards (p. 24).

Nolen, Haladyna, and Haas (1992) found that the perception of pressure due to the accountability system was enough to cause educators to manipulate test scores. In a survey of Arizona educators, the researchers found that more than 43% of teachers believed that administrators and school boards used test scores to evaluated teacher effectiveness. This would be perceived as a great deal of pressure, since only 7% of teachers believed test scores should be used to evaluate teacher effectiveness. When the researchers interviewed administrators, they found that only 15% actually used test scores in the evaluation of teacher performance. The teachers further perceived pressure from

administrators to manipulate test scores. The researchers found that 7% of teachers believed they were encouraged to teach actual test items to their students. Furthermore, more than one-third of teachers believed they were encouraged to use more class time than required for test preparation activities and more than two-thirds believed they were encouraged to teach test-taking skills, focus on skills from the test, and use the item format from the test on classroom tests (pp. 11-12).

Survey results from the National Board on Educational Testing and Public Policy (Pedulla et al., 2003) further supports the notion that a state's accountability systems is the reason why educators manipulate test scores. Although obtaining a response rate of only 35%, the researchers found that 72% of the 4,195 teachers responding to the survey agreed to the statement, "The state-mandated testing programs lead some teachers in my school to teach in ways that contradict their own ideas of good educational practice" (p. 31). Jacob and Levitt (2003) also reached this conclusion, finding that "a high-stakes testing environment increases probability that a teacher would cheat" (p. 17).

In reviewing published news reports of educators manipulating test scores, Nichols and Berliner (2005) concluded that the reports provided evidence of Campbell's Law. Campbell's Law states, "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell, 1975, p. 35). The researchers argued that pressures from state accountability systems lead educators to feel justified in manipulating test scores, stating, "It is plausible that teachers and administrators are trying to resist a system they see as corrupt or unfair, as do tax, religious, and civil rights protestors across this nation" (Nichols & Berliner, 2004, p. 24).

Gay (1990) found further evidence of this feeling of justification. Of the 161 respondents to his survey, Gay found 60% rationalized the use of manipulations in order to improve the image of the teacher and 11% justified manipulations to help students.

Shepard (1990) found that educators believe "a district is at a disadvantage if it plays fair by teaching to a broad curricular domain and by avoiding more than one-time practice on test formats" (p. 21). Cizek (2003) also reached the conclusion that "educators appear to be growing increasingly indifferent toward cheating and even increasingly to feel that cheating is a justifiable response to externally mandated tests" (pp. 375-376).

Educators Unaware of Manipulations & Their Impact

Another reason why educators manipulate test scores is because they are unaware of the definition and impact of manipulations. Educators simply do not know which behaviors are manipulations and why they should not manipulate test scores. Cizek (2003) claims:

> There is an abundance of information to guide test takers and test administrators in how to avoid inappropriate testing practices. For their part, test developers usually produce carefully scripted directions for administering their tests and provide clear guidelines as to which kinds of behaviors on the part of examinees and administrators are permissible and which are not. Acceptable and unacceptable behaviors are sometimes formalized in state administrative codes or statutes.... Numerous professional organizations have published statements on cheating. (pp. 364-365).

To support Cizek's claim, Appendix C displays some of the codes and standards endorsed by the American Counseling Association, the American Educational Research Association, the American Psychological Association, the American Speech-Language-Hearing Association, the Joint Committee on Testing Practices, the National Association of School Psychologists, the National Association of Test Directors, the National Council on Measurement in Education, and the National Education Association. These codes and standards all provide guidance to educators in determining which testing behaviors are appropriate or inappropriate. The existence of these codes and standards led Cizek (2003) to conclude:

> ... there has not generally been a dissemination problem regarding what constitutes integrity in testing or cheating on tests. Virtually

> everyone involved in testing knows how to administer tests that yield credible, accurate results. (p. 365).

Evidence from teacher surveys, however, contradict Cizek's conclusion. Surveys have shown that educators do not agree with administrators or testing specialists as to which behaviors are appropriate or inappropriate in preparing for or administering tests. Kher-Durlabhji and Lacina-Gifford (1992) asked pre-service teachers to determine which testing activities are appropriate or inappropriate to use. Table 2.3 displays the percent of respondents who rated each testing practice as appropriate. While some of the listed activities are appropriate, the fact that some teachers believed changing completed answer sheets or presenting actual test items for practice are appropriate demonstrates that teachers do not understand what behaviors are manipulations.

Lai and Waltman (2007) similarly surveyed a large sample of Iowa public school teachers to determine their perceptions of the ethicality of testing behaviors. Table 2.4 displays the percent of respondents who believed each testing activity was ethical. The researchers found "unexpectedly high percentages of teachers rated practicing with exactly the same test that will be administered this year as being 'very ethical'" (p. 11) and only a median 80% of teachers within schools viewed this practice as being unethical. Another surprising result was that only 75% of teachers believed that "providing instruction without checking the content of the test" was an ethical practice. While some behaviors considered to be ethical may also be manipulations, these results provide further evidence that educators are unaware of which testing behaviors are manipulations.

Lai and Waltman (2007) similarly surveyed a large sample of Iowa public school teachers to determine their perceptions of the ethicality of testing behaviors. Table 2.4 displays the percent of respondents who believed each testing activity was ethical. The table provides further evidence that educators are unaware of which testing behaviors are appropriate and which behaviors are manipulations. The researchers found "unexpectedly high percentages of teachers rated practicing with exactly the same test

that will be administered this year as being 'very ethical'" (p. 11) and only a median 80% of teachers within schools viewed this practice as being unethical. Another surprising result was that only 75% of teachers believed that "providing instruction without checking the content of the test" was an ethical practice. The researchers found that the results of the survey "raise questions about the extent to which teachers understand the testing procedures used in Iowa and are aware of existing professional standards dictating appropriate versus inappropriate testing practices" (p. 49).

Moore (1994) presented 42 elementary school teachers and 10 testing specialists from the Midwest a list of 40 test administration and preparation practices. The subjects were asked to rate the inappropriateness of each activity on a five-point scale (1 = appropriate; 5 = inappropriate). Table 2.5 shows the mean inappropriateness ratings assigned by teachers and testing specialists to eight categories of test administration and preparation practices. The table shows significant differences between teacher and testing specialist inappropriateness ratings in six of the eight categories of practices. Testing specialists rated practices as being more inappropriate and teachers rated the practices as being more appropriate. Popham (1991) conducted a similar analysis and also found discrepancies between teachers' and administrators' perceptions of the relative appropriateness of testing behaviors. These discrepancies in perceived appropriateness once again demonstrate that educators are unaware of which testing behaviors are inappropriate manipulations.

Table 2.3 Percentage of pre-service teachers rating each activity as appropriate

98.6%	Encourage students to do their best
96.0%	Teach test-taking skills
86.5%	Check student's completed answer sheets
79.7%	Send note home to parents to elicit cooperation
66.2%	Use commercial test preparation materials
37.9%	Teach according to test objectives
37.8%	Develop curriculum based on test
37.4%	Practice alternative forms of test
36.5%	Teaching objectives based on standardized test
23.4%	Rephrasing wording of questions
8.1%	Present actual test items for practice
2.7%	Allow more time than allocated for testing
1.4%	Change completed answer sheets
1.4%	Give hints or clues
1.4%	No special test preparation
0%	Dismiss low-achieving students from test taking
0%	Change answers of low-achieving students

Source: Kher-Durlabhji, N. & Lacina-Gifford, L. J. (1992).

Table 2.4 Median percent of Iowa teachers within schools rating each activity as ethical or unethical

Ethical	Behavior	Unethical
100%	Teach test-taking skills	0%
100%	Use previous year's test data to inform instruction	0%
80%	Use practice tests	0%
75%	Provide instruction without checking the content of the test	0%
75%	Structure all/most classroom tests like the ITBS (standardized test)	0%
60%	Review tested content/skills prior to testing	20%
20%	Practice with last year's questions	67%
17%	Routinely provide instruction only on tested content areas	67%
11%	Practice with the same test questions	80%

Source: Lai, E. & Waltman, K. (2007).

Table 2.5 Teacher and testing specialist mean appropriateness ratings

Testing Behavior	Teacher Mean Rating	Testing Specialist Mean Rating
Generalized Test-Taking	1.2	1.1
Motivational Activities	1.6	2.3*
Same Format Preparation	1.7	2.4*
Pretest Intervention	2.4	2.8*
Previous Form Preparation	2.5	3.6*
Posttest Intervention	3.7	3.8
Current Form Preparation	2.9	3.8*
During Test Intervention	2.9	3.7*

* $p < .05$

Source: Moore, W.P. (1994)

Educators do have access to an abundance of information to guide them in selecting testing practices. If dissemination of this information is not the problem, perhaps the reason why educators manipulate test scores is because of an *overabundance* of professional codes and standards. In addition to the codes and standards listed in Appendix C, many states and school districts have rules and laws regarding test security. Additionally, several researchers have provided guidelines for teachers to evaluate the appropriateness of test preparation activities. These guidelines are displayed in Tables 2.6 and 2.7. The overwhelming number of guidelines, codes, standards, rules, and evaluative criteria may overwhelm educators who try to learn which testing behaviors are appropriate or inappropriate. Furthermore, since these guidelines are not requirements, educators may simply ignore them. Finally, educators such as Kilian (1992) have found these guidelines and rules to be, at best, vague, and at worst, contradictory. The lack of clear guidance has led to the lack of understanding of which testing behaviors are inappropriate which may be a major reason why educators manipulate test scores.

Table 2.6 Examples of appropriate and inappropriate test preparation practices

Source	Appropriate	Uncertain	Inappropriate
Frederickson (1984)	1. Physical/emotional/intellectual preparation 2. Time-use and error-avoidance skills 3. Guessing strategies		
Mehrens & Kaminski (1989)	1. General instruction on objectives (no reference to test) 2. Teach test-taking skills	1. Teach objectives from organizations that scan objectives from many tests 2. Teach objectives matching test objectives 3. Use the format of the test questions	1. Teach from published parallel form of test 2. Instruction from the actual test
Haladyna et al. (1991)[a]	1. Training in test-wiseness skills 2. Sanitizing answer sheets (for all students) 3. Increasing motivation via appeals to parents & students		1. Developing curriculum based on test content 2. Preparing objectives based on test items to teach 3. Presenting items similar to test items 4. Using commercial test prep. packages 5. Dismissing low-achieving students from testing[b] 6. Presenting items verbatim from the test[b]
Mehrens et al. (1993)[c]		1. Unauthorized access to test materials 2. Failure to keep materials in locked storage 3. Tampering with sealed test booklets 4. Administration of parallel forms prior to testing 5. Using specific content of test items in instruction 6. Failure to sign and return security forms	1. Practicing with actual test items prior to testing 2. Providing answers or coaching students 3. Altering answer documents prior to scoring 4. Photocopying test materials
Lai & Waltman (2007)[d]	1. Teach test-taking skills 2. Use previous year's data to inform instruction 3. Provide instruction without checking test content 4. Use practice tests 5. Structure all/most classroom tests like the actual test 6. Review tested content/skills prior to testing	1. Practice with last year's questions	1. Practice with the same test questions 2. Provide instruction only on tested content
CEA (2007)[e]	1. Ensure the curriculum is being taught effectively 2. Ensure students are ready physically & psychologically 3. Ensure students are using appropriate time management and deductive reasoning strategies 4. Ensure environments are conducive to performance 5. Ensure test administrators are knowledgeable/prepared 6. Ensure teaching and learning climates in classrooms and schools are positive and productive	1. Provide students a review of tested content 2. Developing classroom tests with items of the same format as the test 3. Special efforts before test administration to teach test-taking skills or increase student motivation for the test	1. Practice with current form of the test 2. Practice with next year's form of the test 3. Practice with previous forms of the test 4. Develop practice tests similar in content/format 5. Drill students for short-term retention 6. Any activity designed primarily to fit the test 7. Activities that must be implemented just before test 8. Teach students to blindly guess/use testing "tricks"

a – the researchers rated the ethicality of practices; skills rated as "unethical" or "highly unethical" are listed as inappropriate in this table.

b – these activities were described as being "highly unethical"

c – actions deemed unacceptable by the vast majority of states are classified as "inappropriate" and actions deemed unacceptable by many but not all states are classified as "uncertain"

d – based on median school-level ethicality rating from Iowa public school teachers

e – these test preparation activities were classified in regards to their use in preparing for the *Iowa Tests of Basic Skills* in Iowa

Table 2.7 Guidelines and criteria to select appropriate test preparation activities

Study	Guidelines / Criteria
Ligon & Jones (1982)	An appropriate test preparation activity "contributes to students' performing near their true achievement levels, and contributes more to their scores than would an equal amount of regular classroom instruction"
Matter (1986)	Inappropriate activities are, "Any additional activities not incorporated into regular ongoing instruction"
Popham (1991)	Test preparation activities should be evaluated through reference to two evaluative standards: Professional ethics: No test preparation activity should violate the ethical standards of the education profession. Practices can be considered unethical if they violate general ethics of theft, cheating, and lying. Practices such as changing student answers on the answer sheet, providing practice items from the actual test, and excluding lower ability students from testing would be examples of violations. Educational defensibility**:** No test preparation practices should increase students' test scores without simultaneously increasing student mastery of the content domain tested. Violations of this standard would include all practices that attempt to improve test performance through test-taking skills, testwiseness skills, motivational strategies, and inappropriate practices such as providing additional examples to students during testing.
Crocker (2006)	Criteria for assessing the appropriateness of proposed preparation activities: Validity, Academic ethics, Fairness, Educational value, Transferability
Iowa Testing Programs (2005)	The appropriateness of any proposed practice should meet either of the two following standards: • It will promote the learning and retention of important knowledge and content skills that students are expected to learn. • It will decrease the chance that students will score lower on the test than they should due to inadequate test-taking skills or limited familiarity with the item formats used on the test. Activities that do not meet one of these criteria are more likely to be unethical, to promote only temporary learning, or to waste instructional time.
CEA (2007)	Test preparation activities must meet three criteria: **Academic Ethics**: The action should not contribute to the misrepresentation or falsification of information The action should not be **perceived** by students, parents, or the community as being dishonest The action should not result in a violation of district policy or copyright (e.g., an illegal act) **Score Meaning & Use**: Test scores should accurately represent student learning related to the specific set of content and skills covered by the test Test scores should not be influenced by a inadequate test-taking skills or limited familiarity with test item formats Test scores should allow users to make accurate inferences related to the larger domain of content and skill areas **Educational Value**: The action should promote learning and long-term retention of content/skills defined by district standards/curriculum The action should provide students with knowledge/skills that apply to a broad range of situations/contexts The amount of instructional time dedicated to test preparation should be warranted in light of educational opportunities lost The actions should be matched with the needs of individual students

Lack of Oversight and Policies

A fourth reason why educators manipulate test scores is because many states have not developed or implemented high-quality policies to prevent these behaviors. Surveys of state departments of education from the National Council on Measurement in Education (Mehrens, Phillips, & Schram, 1993) and Caveon (Sorensen, 2006) along with a survey of Iowa public school districts (Thiessen, 2007) have found that many states and school districts have not implemented simple policies to deter educators from manipulating test scores. Table 2.8 displays the results of these surveys.

Mehrens, Phillips, and Schram (1993) surveyed 46 state departments of education and found the majority claimed to have written test security policies. 65% of states claimed to have written policies addressing test preparation activities, 88% claimed to have policies addressing the security of test materials, and 95% claimed to have policies addressing test administration activities. More than a decade later, after the accountability systems under NCLB had been implemented, Sorensen (2006) conducted a similar survey and found fewer states claimed to have test security policies. While 77% claimed to have policies addressing test preparation activities, only 63% had policies addressing the security of test materials, and less than half (47%) had policies addressing test administration behaviors. It is not known why fewer states claimed to have test security policies in 2006 than in 1993.

Because Iowa does not have a state test security policy, Thiessen (2007) surveyed 154 Iowa public school districts to determine the existence and quality of test security policies at the district level. One year after the Iowa Department of Education disseminated a sample policy and guidance to develop their own policies to school districts, only 27% reported adopting a test security policy. 73% of Iowa public school districts had not yet developed a test security policy and 65% had no plans to adopt a policy in the near future.

Table 2.8 State and school district test security policies

	Mehrens, Phillips, & Schram, 1993	Sorensen, 2006	Thiessen, 2007
Sample	46 state departments of education	34 state departments of education	154 Iowa public school districts
Have written policies addressing:			
Test preparation activities	65%	77%	27%
Test administration activities	95%	47%	27%
Security of test materials	88%	63%	27%
Have no plans to adopt a written policy in the near future	---	---	65%
Policies:			
Require test proctors to be trained	---	78%	78%
Require independent monitoring of test administration	---	---	31%
Identify individuals responsible for responding to incidents	---	65%	90%
Provide a separate budget for test security	---	5%	---
Have policies to specify that:			
Test materials must be sealed before administration	48%	---	---
New test forms must be used each year	54%	---	---
Teachers cannot examine tests before administration	62%	---	---
Routinely run statistical analyses of answer sheets to check for:	---	52%	6%
Unusual number or pattern of erasures	49%	---	---
Unusual score fluctuations	64%	---	---
If cheating or manipulations are suspected:			
The incident is investigated	64%	81%	---
A written policy guides the investigation	24%	47%	---
Percentage of suspected incidents that are later confirmed	50%	---	---
If cheating or manipulations are confirmed:			
No sanctions are imposed (or only a letter is sent)	52%	11% - 18%	---
The guilty party is given a stern warning or reprimand	19%	61%	---
The guilty party is suspended	12%	34%	---
The guilty party is dismissed	8%	45%	---

The policies that are adopted by states and school districts are not all of high quality. In fact, some policies ignore fundamental methods to prevent test score manipulations. For example, Sorensen (2006) and Thiessen (2007) both found that only 78% of states and Iowa school districts required their test proctors to be trained prior to test administration. Survey results also show that less than two-thirds of states have

identified an individual or group of individuals to be responsible for answering questions about test security or responding to test score manipulations, and only 5% of states provide a separate budget for test security. In Iowa, less than one-third of the school districts surveyed had policies requiring independent monitoring of the test administration. Of those districts that did have policies, Thiessen found that only 8% had policies that could be effective in deterring test score manipulations.

Mehrens, Phillips, and Schram (1993) found that many states did not require basic test materials security precautions. They found that less than two-thirds of state policies specify that teachers cannot examine test questions before the test administration and only 54% of state policies require new test questions to be used annually. They also found that less than half of the state policies require test materials to be sealed prior to administration.

Even fewer states had policies that called for routine statistical analyses of answer sheets to detect potential manipulations. 49% of states reported routinely analyzing answer sheets for unusual patterns of erasures and 64% reported routinely checking for unusual score gains in the 1993 survey (Mehrens, Phillips, & Schram, 1993). A 2006 poll conducted by the Philadelphia Inquirer found that fewer than half of all states attempt to detect manipulations on their state tests (Patrick & Eichel, 2006). The 2006 survey from Caveon found that 52% of states claimed to routinely run statistical analyses to check for evidence of manipulations (Sorensen, 2006). This survey also found that 25% of states had no plans to implement statistical analysis methods to detect manipulations. Thiessen (2007) found that only 6% of Iowa public school districts routinely conducted these analyses.

The surveys of state departments of education also found that many state policies are weak when it comes to investigating reports of manipulation incidents. Mehrens, Phillips, and Schram (1993) found that when educators are suspected of cheating on state tests, only 64% of those incidents are investigated. The researchers also found that when

states do decide to investigate reported incidents, only 24% of states have a written policy to guide the investigation. Sorensen (2006) found that the situation may have improved slightly since 1993, finding that 81% of states claim to investigate suspected incidents of cheating and 47% have a written policy that prescribes actions to be taken when cheating is suspected.

According to the 1993 survey, half of all suspected incidents of cheating are later confirmed (Mehrens, Phillips, & Schram, 1993). The states admit, however, that the sanctions imposed on confirmed manipulators are not standardized. The 1993 survey found that in more than half of all confirmed incidents; the guilty party was either not penalized at all or only sent a letter from the state department of education. By 2006, only 11% of states did not penalize confirmed cheaters (Sorensen, 2006). In 1993, a stern warning or reprimand was given in 19% of confirmed cases; the guilty party was suspended in 12% of confirmed cases; and the guilty party was dismissed in 8% of confirmed cases of educator cheating. By 2006, the sanctions increased in magnitude, with 61% of confirmed cheaters receiving a stern reprimand, 34% receiving a suspension, and 45% being dismissed.

The lack of policies providing basic considerations of test materials security, test preparation activities, and test administration behaviors might not encourage educators to manipulate test scores, but it certainly does nothing to deter or prevent educators from engaging in these behaviors. As will be discussed in the next section, high quality, enforceable policies can deter educators from manipulating test scores.

How to Prevent Manipulations: Evaluation of State Policies

Researchers have provided several suggestions to deter or prevent educators from manipulating test scores. In their research on student cheating, Aiken (1991), Burns (1988), Cizek (1999), and Singhal and Johnson (1983) suggested that test developers can prevent manipulations by modifying the tests used to make high-stakes decisions. The

researchers recommend test publishers develop constructed-response test items, suggesting that it would be more difficult for educators to manipulate scores from these constructed-response tests than it would be for multiple-choice tests. The researchers also suggested that test developers develop new test forms with new items each time the test is administered to make it more difficult for educators to manipulate scores.

Impara and Foster (2006) suggest that while these methods may be effective in preventing students from cheating on tests, "item and test development strategies do little to reduce" educator score manipulations (p. 93). Also, developing new forms of constructed-response tests for each administration would be labor-intensive and inefficient. Test developers should focus on developing the best items to measure the construct of interest; not developing items that are most resistant to manipulations.

Another method to prevent test score manipulations might be to catch and punish educators who manipulate test scores. This would require states to implement methods to detect manipulations, such as statistical analyses of answer sheets or surveys after test administration, and to strengthen the sanctions imposed upon educators found to manipulate scores.

Ignoring the facts that statistical analyses have been shown to be ineffective in detecting manipulations (Chason & Maller, 1996; Impara, Kingsbury, Maynes, & Fitzgerald, 2005; Iwamoto, Ningester, & Luecht, 1996) and that states have been reluctant to punish educators who manipulate test scores, suppose states could accurately detect manipulators. Even if states punished these manipulators, this *after-the-fact* approach to deter manipulations would be labor-intensive and, if used as the only deterrent, would most likely be ineffective. In their study on student cheating, Bunn, Caudill, and Gropper (1992) found that both the expectation and severity of punishment had no effect on reducing cheating behaviors in students.

A third method to reduce the number of educators who manipulate test scores would be by developing, implementing, and disseminating high quality policies that both

discourage manipulations and encourage honesty and integrity. In their study on student cheating, McCabe and Trevino (1993) found that students were less likely to cheat if their schools had severe penalties for cheating coupled with high quality policies or honor codes on student cheating. Based on a decade of research from more than 14,000 students, the researchers found that neither sanctions nor honor codes alone reduced cheating, but that the combination of the two was effective at reducing student cheating by as much as 20% (McCabe & Trevino, 2002). In order to be effective, the honor code must explain why academic integrity is important, describe which behaviors are appropriate or inappropriate, and clearly show the school's commitment to academic integrity. The sanctions described in the policy must be significant and consistently applied to those caught cheating. The researchers also found that in order to be effective, the policies must be developed with input from students and supported by top administrators.

Cizek (2003) suggested that the combination of policies and sanctions might also work to deter educators from manipulating test scores. While educators have had an overabundance of professional codes and standards and test administration manuals to guide their behavior, these guidelines have not been enforceable and educators have not been held accountable for following them. Policies developed by state boards of education, on the other hand, may be effective because educators would be required to follow them. Cizek (1999, 2001) recommended that states bear responsibility for developing policies to prevent educators from manipulating test scores.

Some states have developed specific policies and regulations to address test score manipulations, but many others have left this task up to individual school districts (Cizek, 1999; Mehrens, Phillips, & Schram, 1993; Patrick & Eichel, 2006). Unfortunately, very little research has been conducted to determine the existence or evaluate the quality of these state and district developed policies. In 1999, Cizek wrote, "Only one study has been conducted to investigate the existence of policies at the elementary and secondary

school level" (1999, p. 171); "Unfortunately, no research has actually examined the content of cheating policies" (p. 174); and that it was not even known if schools, school districts, or states had any policies to address educator manipulations (p. 171). Since that time, Thiessen (2007) conducted an evaluation of test security policies in Iowa public school districts and found that 73% of districts had no policy, 22% had policies that were inadequate to deter test score manipulations, and 5% had policies that could effectively prevent test score manipulations.

Due to a lack of useful frameworks, the relationship between test security policies and test score manipulations is not currently known. To address this, a framework is developed to evaluate the quality of state test security policies. Then, analyses are conducted to determine the relationship between the quality of those policies and estimates of test score manipulations.

Test Security Policy Content: Evaluative Framework

As Cizek (1999) noted, little research exists to evaluate the content of state policies to deter educators from manipulating test scores. In their surveys of state departments of education, Mehrens, Phillips, and Schram (1993) and Sorensen (2006) provides general guidelines such as recommending states conduct statistical analyses of answer sheets and outline sanctions for those caught manipulating scores. In his discussion of policies and honor codes, Cizek (1999, 2001) recommended that states describe specific activities in defining what testing behaviors are appropriate or inappropriate. Professional codes and standards, example policies from test publishers (Harcourt Assessment, 2006, 2007; Iowa Testing Programs, 2005; Riverside, 2006), and state departments of education also provide guidance as to the content of effective test security policies. Finally, Thiessen (2007) provided content recommendations in an evaluative framework for the development, adoption, and implementation of district test security policies in Iowa. These sources all provide guidance as to what content a test

security policies needs to effectively prevent test score inflation by deterring educators from manipulating the teaching philosophy or process, examinee pool, test administration, or score reports or standards. The recommendations, along with the specific manipulations they are intended to deter, are summarized in Table 2.9.

If state test security policies are to foil educators from manipulating test scores, these policies must:

- (F) Formalize beliefs of state educators regarding the role of testing and practices
- (O) Oversee test preparation, administration, and scoring activities
- (I) Inform educators about why some behaviors and activities are unacceptable
- (L) Limit opportunities for educators to manipulate test scores

The policy content recommendations can be reclassified to fit into this framework to FOIL test score manipulations. Table 2.10 displays the recommendations under this FOIL framework.

The first way in which a state test security policy can foil test score manipulations is by formalizing both state educators' beliefs about the role of testing in education and current state testing practices. This formalization begins with state educators providing input into the content of the written test security policy. In their review of research on student cheating, McCabe and Trevino (2002) found that school cheating policies were more effective if students were encouraged to assist in the development of the policy content. Similar results could be found if educators assist in the development of state test security policies.

The formalization of a state test security policy also requires states to disseminate their policies and ensure all educators understand the policies. Cizek (2003) made a similar recommendation after examining several published news reports on test score manipulations:

> Reports of cheating are often accompanied by protestations from the guilty parties that they did nothing wrong. Every implementation of high-stakes tests should be accompanied by

> dissemination of clear guidelines regarding permissible and impermissible behaviors. Such reminders should be clearly worded, pilot tested, distributed, and signed by all who handle testing materials, including test site supervisors, proctors, and examinees (pp. 377-378).

If educators do not understand the content, the policy should identify individuals responsible for answering questions about policy content. Cizek (1999) recommended identifying an individual in charge to prevent student cheating and test publisher Harcourt Assessment (2007) requires the identification of such an individual before shipping tests to a customer.

An effective test security policy should also formalize the state's current testing practices. The policy should outline procedures for handling testing materials and testing irregularities, for re-testing students, and for correcting possible scoring errors. The policy should also formalize the state's belief that test score manipulations are unacceptable. To do this, the policy should require mandatory reporting of all incidents of manipulations, while, at the same time, providing protection for those reporting suspected incidents. These recommendations are based on the finding of Gay (1990) that only 20% of educators were willing to report testing irregularities to their school administrators (p. 4). Based on research into effective policies to prevent student cheating (McCabe & Trevino, 2002), state test security policies should outline the procedures that will be used to investigate suspected cases of manipulations and specify the sanctions imposed on those found guilty of manipulating test scores. Cizek (2003) made similar recommendations, stating:

> Enforce penalties for cheating and change the system of investigation. …there are strong disincentives for educational personnel to report cheating; and in most jurisdictions, the responsibility for investigating cheating involves personnel at the school or district level and agencies such as boards of education with an inherent conflict of interest when it comes to ferreting out inappropriately high apparent student achievement. Revised procedures should include… increased protection for whistleblowers; more streamlined procedures and stiffer penalties for cheating, including permanent disqualification from teaching within a state and more coordinated sharing of information regarding educators who have had their licenses revoked; and

> delegation of responsibility for investigating incidents of cheating to an independent agency (p. 381).

The next way in which a state test security policy can foil test score manipulations is by overseeing all aspects of test security. As recommended by the test security firm Caveon (Sorensen, 2006) and Cizek (2003), states should regularly audit the security of their current testing programs. Regular audits serve to evaluate the effectiveness of a state's test security policies and encourage states to refocus efforts on test security.

Effective state policies should also provide for oversight of the administration of tests. As Cizek notes, many tests are administered behind closed doors with little independent oversight (2003, p. 381). Jacob and Levitt (2003) found that teachers who administered exams to their own students without independent oversight were 50% more likely to cheat. By requiring independent monitoring of the test administration in a random sample of classrooms, state test security policies can ensure that test administration directions are followed and that educators will not give answers or hints to students, provide students with more time to complete the test, or provide students with inappropriate reference materials or tools.

Effective policies should also provide for oversight in the form of statistical analyses of test scores and answer sheets. State policies should require all student answer sheets to be analyzed for unusual patterns of erasures (Qualls, 2001), unusual patterns of responses (Cizek, 2003; Jacob & Levitt, 2003, 2004; Sorensen, 2006; Wesolowsky, 1990), and unusual score fluctuations (Cizek, 1999; Jacob & Levitt, 2003). Bellezza and Bellezza (1989) found that when examinees were made aware that statistical analyses would be used to identify cheaters, the incidence of cheating declined from 5% to 1%. The provision for statistical analyses of student answer sheets could have a similar effect of reducing educator manipulations of test scores.

Table 2.9 Content Recommendations for Test Security Policies

	Policy Recommendations
General	Educators should have input into the development of policy content (Cizek, 1999, 2003; McCabe & Trevino, 2002) Policy content should be clearly worded and signatures should be obtained to ensure information is disseminated and understood (Cizek, 1999, 2003; McCabe & Trevino, 2002) The policy should identify individuals who are in charge of answering questions about policy content (Cizek, 1999; Harcourt Assessment, 2007) The policy should require mandatory reporting of all incidents of manipulations (Cizek, 2003; Gay, 1990) The policy should provide protections for those reporting incidents of manipulations (Cizek, 2003) The policy should outline due process and procedures to investigate and handle incidents of suspected manipulations. Ideally, the policy should call for an independent agency to handle investigations. (Cizek, 1999; McCabe & Trevino, 2002) The policy should outline the sanctions imposed on those found to have manipulated test scores (McCabe & Trevino, 2002) The policy should provide for regular audits of test security (Cizek, 2003; Sorensen, 2006)
Manipulate Teaching Philosophy or Process Making copies of test prior to test administration	The policy should explain copyright laws and penalties for violating copyright laws (Harcourt Assessment, 2007; Iowa Testing Programs, 2005; Riverside Publishing, 2006)

Table 2.9 Continued

	The policy should explain copyright laws and penalties for violating copyright laws (Harcourt Assessment, 2007; Iowa Testing Programs, 2005; Riverside Publishing, 2006) The policy should specify who has access to test materials and how to document who handles test materials (Cizek, 2003; Sorensen, 2006) The policy should require test materials to be sealed prior to test administration (Cizek, 1999, 2003; Harcourt Assessment, 2007; Iowa Testing Programs, 2005; Sorensen, 2006) The policy should limit the amount of time educators have access to tests before and after administration (Cizek, 2003; Sorensen, 2006)
Practicing with items identical or similar to the test Practice with last year's (alternate form) test items Practice with items of the same format as the test Use commercial test preparation packages Teaching test-taking skills; test-wiseness Teaching content from specific test items Focusing resources on students closest to proficiency Teaching only content found on the test Changing curricula to better match the test	The policy should require the use of multiple test forms. Ideally, new test forms would be administered each year (Cizek, 2003; Crocker, 2003, 2006) The policy should provide examples of specific appropriate and inappropriate test preparation activities (Cizek, 2003; Moore, 1994; Popham, 1991; McCabe & Trevino, 2002) The policy should provide guidelines as to how much instructional time should be spent on test preparation activities (Lai & Waltman, 2007; Moore, 1994; Popham, 1991) The policy should explain the importance of validity and test scores generalizing to a broader domain (Lai & Waltman, 2007; McCabe & Trevino, 2002) Explain the uses of test scores beyond accountability (for example. to make instructional improvements) (McCabe & Trevino, 2002)
Manipulate Examinee Pool Excluding students from testing Having high-scoring students take test multiple times	No Child Left Behind test participation requirements should be explained, including the testing of disabled students and English language learners. (Cullen & Reback, 2006; Figlio & Getzler, 2002, Jacob, 2007)

Table 2.9 Continued

Providing inappropriate special education placement Bribing or paying students to increase test scores Increasing caloric content of school meals	The policy should explain why accommodations are used in test administration and provide examples of appropriate and inappropriate testing accommodations (or make reference to materials that provide these examples). (McCabe & Trevino, 2002) The policy should provide examples of appropriate and inappropriate school- or classroom-level activities on the day of testing. (McCabe & Trevino, 2002)
Manipulate Test Administration Altering a student's answer sheet (changing answers) Giving students answers Checking or pointing out incorrect answers Giving students (non)verbal hints on test items Not following test administration procedures exactly Allowing students to work together during testing Ignoring students who are cheating Giving students additional examples Providing students extra time Rephrasing test items for students Reading items that are to be read by students Answering questions about test content Providing students with reference materials or tools Having students fill-in unanswered items Providing inappropriate accommodations to students Sanitizing answer sheets (cleaning before scoring) Review skills that will be on tomorrow's test	The policy should provide for monitoring of the test administration. Ideally, independent monitors would be used to oversee test administration in a randomly selected sample of classrooms or schools. (Cizek, 2003) The policy should require all test proctors to be trained. (Lai & Waltman, 2003; Cizek, 2003) The policy should provide specific examples of appropriate and inappropriate test administration behaviors (or make reference to materials that provide examples), including how to respond to student questions and what materials are allowed during testing. (Cizek, 2003; McCabe & Trevino, 2002) The policy should explain the importance of standardization and following test administration procedures. (McCabe & Trevino, 2002) The policy should provide for statistical analyses of answer sheets to check for unusual erasure patterns, response patterns, or score fluctuations. (Bellezza & Bellezza, 1989; Cizek, 2003; Jacob & Levitt, 2003, 2004; Qualls, 2001; Sorensen, 2006; Wesolowsky, 1990) The policy should provide specific examples of how to handle testing irregularities, including how to clean student answer sheets following testing (Cizek, 2003; McCabe & Trevino, 2002)

Table 2.9 Continued

Manipulate Score Reports or Standards	
	The policy should outline procedures to be followed if test scores are suspected of being incorrect, including procedures for re-testing. (Cizek, 2003)
Changing student test scores on official records	
Providing false IDs so scores won't count	The policy should provide a system (barcodes, for example) to ensure accurate student information is matched to test scores (Cizek, 2003)
Misrepresenting data	The policy should provide examples of appropriate and inappropriate interpretations and uses of test scores
Changing criteria for proficiency; making test easier	(McCabe & Trevino, 2002)

Table 2.10 FOIL Framework for Evaluating Test Security Policy Content

Formalize beliefs of state educators regarding the role of testing and practices • Educators should have input into the development of the content of the written state policy • Policy content should be clearly worded and signatures should be obtained to ensure information is disseminated and understood • The policy should identify individuals who are in charge of answering questions about policy content • The policy should require mandatory reporting of all incidents of manipulations • The policy should provide protections for those reporting incidents of manipulations • The policy should outline due process and procedures to investigate and handle incidents of suspected manipulations • The policy should outline the sanctions imposed on those found to have manipulated test scores • The policy should outline procedures to be followed if test scores are suspected of being incorrect, including procedures for re-testing • The policy should provide a system (barcodes, for example) to ensure accurate student information is matched to test scores • The policy should provide specific examples of how to handle testing irregularities, including how to clean student answer sheets following testing
Oversee test preparation, administration, and scoring activities • The policy should provide for regular audits of test security • The policy should provide for monitoring of the test administration. Ideally, independent monitors would be used to oversee test administration in a randomly selected sample of classrooms or schools • The policy should provide for statistical analyses of answer sheets to check for unusual erasure patterns, response patterns, or score fluctuations.
Inform educators about why some behaviors and activities are unacceptable • The policy should explain copyright laws and penalties for violating copyright laws • The policy should provide examples of specific appropriate and inappropriate test preparation activities • The policy should provide guidance as to how much time should be spent on test preparation activities • The policy should explain the importance of validity and test scores generalizing to a broader domain • Explain the uses of test scores beyond accountability (for example. to make instructional improvements) • No Child Left Behind test participation requirements should be explained, including the testing of disabled students and English language learners • The policy should explain why accommodations are used in test administration and provide examples of appropriate and inappropriate accommodations (or make reference to materials that provide these examples) • The policy should provide examples of appropriate and inappropriate school- or classroom-level activities on the day of testing • The policy should require all test proctors to be trained • The policy should provide specific examples of appropriate and inappropriate test administration behaviors (or make reference to materials that provide examples), including how to respond to student questions and what materials are allowed during testing • The policy should explain the importance of standardization and following test administration procedures • The policy should provide examples of appropriate and inappropriate interpretations and uses of test scores
Limit opportunities for educators to manipulate test scores • The policy should specify who has access to test materials and how to document who handles test materials • The policy should require test materials to be sealed prior to test administration • The policy should limit the amount of time educators have access to tests before and after administration • The policy should require the use of multiple test forms. Ideally, new test forms would be administered each year

The third way in which state test security policies can deter manipulations is by informing educators as to what behaviors are acceptable and why other behaviors or activities are unacceptable. Recall that a major reason why educators manipulate test scores is that they are simply unaware as to what behaviors or activities constitute manipulations. Policies that explain copyright laws, NCLB requirements, the role of test preparation, the purposes of testing, the uses of test scores, the importance of validity, and the importance of standardized test administration procedures may help reduce test score manipulations.

McCabe and Trevino (1993, 2002) found that when schools informed students about the importance of testing and the seriousness of cheating through honor codes, student cheating reduced. They found that, in order to be effective, these honor codes must also provide specific examples of appropriate and inappropriate behaviors. Likewise, state test security policies should provide specific examples of appropriate and inappropriate test preparation, administration, and scoring behaviors and activities. If state policies were developed with input from educators, then these policies would represent state educators' collective beliefs as to the appropriateness of each testing activity. By codifying these collective beliefs, the state policies would be more effective in reducing manipulations than the overabundance of professional guidelines, standards, and codes currently available to educators.

State policies should also inform educators by requiring all test proctors to be trained regularly. Cizek (2003) noted:

> Too often, the qualifications for proctoring exams are only faintly spelled out, the training provided is minimal if any, and no incentives exist to heighten proctors' vigilance or pursuit of instances of cheating. Proper training must include instruction on methods examinees use to cheat and effective procedures for documenting on-site testing irregularities (pp. 380-381).

Training would ensure that educators are aware of which behaviors are unacceptable and the sanctions they will face if they manipulate test scores. This combination of

information about the importance of testing, specific examples of appropriate and inappropriate activities, and sanctions was found to be effective in reducing student cheating at the collegiate level (McCabe, Trevino, Butterfield, 2001).

The fourth way in which state test security policies can deter manipulations is by limiting educators' opportunities to manipulate scores. This can be done by limiting access to test materials and by administering new test forms annually. As many test publishers recommend or require (Harcourt Assessment, 2007; Iowa Testing Programs, 2005; Riverside Publishing, 2006) test materials should be sealed prior to administration. Policies should also outline the handling of test materials to limit the amount of time educators have access to materials before and after the test administration. Sorensen (2006) and Cizek (2003) both recommend that states specify who has access to materials and document anyone who has been given access. This would deter educators from manipulating the teaching process or philosophy through inappropriate test administration and from manipulating the test administration by changing student answers. Cizek (2003) further recommends that states administer new test forms annually (or that test disclosure laws be revised) in order to prevent educators from manipulating test scores through inappropriate test preparation activities.

Relationship Between Test Security Policies and Score Trend Discrepancies

In order to determine if test security policies are effective in deterring educators from manipulating test scores, the impact of test score manipulations must first be estimated. Unfortunately, it is extremely difficult to do this. While the retesting experiments conducted in Chicago (Jacob and Levitt, 2003, 2004; Perlman, 1985) provide evidence of the impact of manipulations on test scores for individual students or classrooms, they do not provide evidence that manipulations significantly impact test scores on a district or state level.

In order to even begin to address the impact of test score manipulations at the state-level, the definition of *manipulation* must again be considered. Recall that the term *manipulation* is defined as any practice used by educators to increase test scores without an equal, corresponding increase in student performance on the underlying construct. This suggests that in order to provide evidence of the impact manipulations, state test scores can be compared to a different measure of the same construct, assuming scores from the two tests scores can be compared on the same scale. If scores from the state test align with scores from the other measure, then one could conclude that manipulations have very little impact on state test results. If, on the other hand, a large discrepancy exists between state test scores and scores on the other measure, then one possible explanation for this discrepancy could be that educators manipulated state test scores.

As will be discussed later, this conclusion would be only one of many possible explanations for the discrepancy between two measures of the same construct. When a large discrepancy exists between two measures of the same construct, all alternative plausible rival hypotheses should be ruled-out before drawing a conclusion of causality (Koretz, 1991). This study will only attempt to find if a relationship exists between the quality of a state's test security policy and discrepancies in score trends as measured by two tests.

Due to the requirements of NCLB, every state has already implemented an accountability system using a state test. These state tests all measure student performance in, at least, reading and mathematics and provide a percentage of students who score at or above a proficient level in each subject. In order to attempt to estimate the impact of manipulations on state level test scores, researchers must choose an appropriate second measure of the same constructs of reading and mathematics. In addition to measuring the same construct, this second measure must be designed, administered, and scored so that its scores cannot be manipulated in the same way as state

test scores. An obvious choice for this second measure would seem to be the National Assessment of Educational Progress (NAEP).

NAEP is a congressionally mandated assessment administered by the National Center for Education Statistics (NCES) (Chadwick, 2006). While states and school districts are required to participate in the testing, not all students are tested on the NAEP each year. As part of the State NAEP program, representative samples of students from grades 4 and 8 are selected from each state to take the test. Instead of testing each student in reading and mathematics, each student is administered a portion of the entire test. These results are then combined to provide average scale scores and percentages of students meeting basic, proficient, and advanced performance standards at the state level.

NAEP is designed and administered in a way that has made it potentially more robust against educator manipulations. This means that it can provide a good comparison to the results from state tests that can be subject to manipulations. First, results from the NAEP are not used to make high-stakes decisions regarding the performance of an individual educator, school, or school district. Because of this, educators should feel no pressure to manipulate test scores on the NAEP. Second, although some items are publicly released after testing, educators do not have access to the items on the NAEP before it is administered. This virtually eliminates the possibility that educators will manipulate the teaching process or philosophy through inappropriate practice or coaching to inflate NAEP scores. Third, by forcing make-up testing for classrooms with less than 90% attendance and by comparing sample demographics to state demographics, the NAEP provides some level of protection against manipulations of the examinee pool (NCES, 2007). Finally, because the U.S. Department of Education hires staff to administer the NAEP and classroom teachers can monitor the administration, educators would have difficulty manipulating the test administration in order to inflate NAEP scores (Massachusetts Department of Education, 2007). Thus, the NAEP provides results that can be more robust against educator manipulation.

Single-Year Comparisons of State and NAEP Results

Because manipulations may have a smaller impact on NAEP scores than state test scores, researchers have developed methods to compare the results from state and NAEP tests. One simple method involves researchers making single-year comparisons of state and NAEP test proficiency results. Studies that have used this method include research from the Civil Rights Project at Harvard University (Lee, 2006), The Education Trust (Hall & Kennedy, 2006), and The Hoover Institution (Peterson & Hess, 2005, 2006).

The logic behind these studies is this: (a) if the state tests and NAEP measure the same constructs of reading and math, and (b) if the state tests and NAEP both define *proficiency* in these constructs, and (c) if the definitions of proficiency are similar, then (d) the percentages of proficient students provided by both tests should be similar. If the reported proficiency rates from the two tests do not provide similar results in a given year, then that discrepancy provides possible evidence that the state tests, which are more susceptible to manipulation, may have been inflated through manipulation. Again, these conclusions are based on strong assumptions that will be discussed later.

Table 2.11 displays an example of the results from these single-year analyses. The second column of Table 2.11 displays the percentage of 8th grade students who scored at or above proficient in mathematics during the 2005 administration of the state test. The third column shows the percentage scoring at or above proficient on the 8th grade NAEP test. As the table shows, the proficiency rates obtained from the state tests are higher than proficiency rates from the NAEP for 46 states. The median state percentage of students scoring at or above a proficient level on state tests was 62% in 2005. The state median percentage of students scoring proficient or above on the NAEP was 30%. Therefore, the median proficiency rate reported from state tests is 2.07 times larger than the median proficiency rate reported from the NAEP. The fifth column of the table displays this information for each state.

Table 2.11 Results from 2005 state and NAEP tests of 8th grade mathematics

	% at or above **proficient** on...		% at or above **basic** on...	Ratio of % proficient or above on **state test** to...	
State	State Test	NAEP	NAEP	% at or above **proficient** on **NAEP**	% at or above **basic** on **NAEP**
Alabama	63	15	53	**4.20**	***1.19***
Tennessee	87	21	61	**4.14**	***1.43***
West Virginia	71	18	60	**3.94**	***1.18***
Mississippi	53	14	52	**3.79**	***1.02***
Oklahoma	69	21	63	**3.29**	***1.10***
Louisiana	51	16	59	**3.19**	**0.86**
Georgia	69	23	62	**3.00**	***1.11***
North Carolina	84	32	72	**2.63**	***1.17***
Virginia	81	33	75	**2.45**	***1.08***
Utah	73	30	71	**2.43**	***1.03***
Arizona	63	26	64	**2.42**	**0.98**
Indiana	71	30	74	**2.37**	**0.96**
Colorado	75	32	70	**2.34**	***1.07***
Idaho	70	30	73	**2.33**	**0.96**
Nevada	49	21	60	**2.33**	**0.82**
Florida	59	26	65	**2.27**	**0.91**
Iowa	74	34	75	**2.18**	**0.99**
Connecticut	76	35	70	**2.17**	***1.09***
Alaska	62	29	69	**2.14**	**0.90**
Michigan	62	29	68	**2.14**	**0.91**
Nebraska	72	35	75	**2.06**	**0.96**
Wisconsin	73	36	72	**2.03**	**0.88**
Pennsylvania	63	31	76	**2.03**	**0.96**
Kansas	68	34	77	**2.00**	**0.88**
Texas	61	31	72	**1.97**	**0.85**
South Dakota	69	36	80	**1.92**	**0.86**
Ohio	63	33	74	**1.91**	**0.85**
Oregon	64	34	72	**1.88**	**0.89**
North Dakota	65	35	68	**1.86**	**0.79**
Illinois	54	29	81	**1.86**	**0.80**
New York	56	31	70	**1.81**	**0.80**
Minnesota	76	43	79	**1.77**	**0.96**
Delaware	53	30	72	**1.77**	**0.74**
Montana	63	36	80	**1.75**	**0.79**
Maryland	52	30	66	**1.73**	**0.79**
New Jersey	62	36	74	**1.72**	**0.84**
New Mexico	24	14	53	**1.71**	**0.45**
California	37	22	57	**1.68**	**0.65**
Rhode Island	39	24	63	**1.63**	**0.62**
New Hampshire	56	35	77	**1.60**	**0.73**
Vermont	60	38	78	**1.58**	**0.77**
Kentucky	36	23	64	**1.57**	**0.56**
Arkansas	33	22	64	**1.50**	**0.52**
Washington	51	36	75	**1.42**	**0.68**
Wyoming	38	29	76	**1.31**	**0.50**
Hawaii	20	18	56	**1.11**	**0.36**
Maine	29	30	74	***0.97***	**0.39**
Massachusetts	39	43	80	***0.91***	**0.49**
South Carolina	23	30	71	***0.77***	**0.32**
Missouri	16	26	68	***0.62***	**0.24**
Median	**62**	**30**	**71**	**2.07**	**0.42**

Source: Hall, D., & Kennedy, S. (2006)

The table shows 46 states reported larger proficiency rates than what were reported by the NAEP. Alabama reported the greatest discrepancy, with a state test proficiency rate 4.20 times larger than the NAEP proficiency rate. Only four states reported a discrepancy in the opposite direction, with Missouri's state test proficiency rate being only 0.62 times the proficiency rate reported by the NAEP.

Based on similar results from state and NAEP testing in 4th and 8th grade reading and mathematics in 2003 and 2005, researchers have concluded that state test scores are inflated. Reports from The Brookings Institution (Ravitch, 2005), The Civil Rights Project (Lee, 2006), The Education Trust (Hall and Kennedy, 2006), and The Hoover Institution (Peterson and Hess, 2006) all conclude that the single-year discrepancies in proficiency rates between state tests and NAEP are due to states manipulating their score standards. Ravitch (2005) concluded that states lower their proficiency standards "for fear of alienating the public and embarrassing public officials responsible for education" (p. 2). Peterson and Hess (2005, 2006), in using similar data to rate each state's accountability system, concluded that state were "tempted to race to the bottom, lowering expectations to ever lower levels so that fewer schools are identified as failing, even when no gains are being made" (p. 1). Lee (2006) found a positive correlation between the strength of a states high-stakes accountability system and the size of the discrepancy in proficiency rates, concluding that states make tests easier and "water down [their] own performance standards" (p. 51) in order to inflate test scores. Hall and Kennedy (2006) reached a similar conclusion, stating, "most state standards for proficiency are closer to the *basic* level on the NAEP" (p. 19) than they are to the NAEP proficiency level.

Other researchers argue that state proficiency standards are closer to the NAEP basic standards because the tests use different definitions of *proficiency*. In a 2007 report, Idaho's NAEP State Coordinator reviewed the literature and developed six guidelines on the proper use of NAEP scores in confirming results from state tests (Stoneberg, 2007ab). Among the guidelines, Stoneberg noted that state and NAEP

definitions of *proficient* were not the same and that "NAEP's percentage at or above *basic* is the most directly comparable statistic for confirming state results" (2007a, p. 7). Stoneberg noted that under NCLB, the U.S. Department of Education required states to define proficiency in terms of grade-level expectations (p. 3). A student scoring proficient according to state standards should represent a student who is achieving at or above grade-level expectations. Stoneberg notes that NAEP, on the other hand, does not consider grade-level expectations in defining proficiency. The National Assessment Governing Board printed the following in a booklet designed to inform the public about interpretations of NAEP scores:

> Achievement levels define performance, not students. Notice that there is no mention of "at grade level" performance in these achievement goals. In particular, it is important to understand clearly that the Proficient achievement level does not refer to "at grade" performance. Nor is performance at the Proficient level synonymous with "proficiency" in the subject. That is, students who may be considered proficient in a subject, given the common usage of the term, might not satisfy the requirements for performance at the NAEP proficiency level. Further, Basic achievement is more than minimal competency. Basic achievement is less than mastery but more than the lowest level of performance on NAEP. (Loomis & Bourque, 2001).

Further supporting the argument that proficiency rates from state tests should not be compared to proficiency rates from the NAEP, the 2004 NAEP Validity Studies Panel (Mosquin & Chromy, 2004) recommended that "of the various statistics that might be used for measuring a gap on the NAEP scale – proportion at or above the basic, proficient, or advanced achievement level, or mean standardized score – the proportion at or above the *basic* achievement level will both have the greatest correlation with the adequate yearly progress statistic and also be the most directly comparable" (p. 12).

Table 2.11 illustrates the impact of comparing state proficiency to the NAEP basic level instead of the NAEP proficient level. The fifth column shows the ratio of the percentage of students scoring proficient on the state test to the percent of students scoring at or above the basic level of achievement on the NAEP. Whereas 46 states had

proficiency rates higher than NAEP proficiency rates, only 11 states had proficiency rates higher than NAEP basic rates. The number of students scoring proficient on the Tennessee state test, for example, was 1.43 times larger than the number of students scoring at a basic level on the NAEP. Missouri, on the other hand, had a NAEP basic rate 4.25 times larger than the state reported proficiency rate of 16%, implying that state standards have been set higher than NAEP standards.

Due to the fact that conclusions from these studies on the impact of manipulations on state test scores differ depending on which NAEP standard (basic or proficient) is used, these single-year state and NAEP comparisons are limited. In an attempt to address this limitation, researchers from the U.S. Department of Education (USDE) (Braun & Qian, 2007), the American Institutes of Research (AIR) (McLaughlin et al., 2000, 2002) have developed another group of single-year comparison methods to compare state test and NAEP results. These methods, which were developed through a series of 12 studies beginning in 1993, involve linking state scores or proficiency standards onto the NAEP score scale (Buckley, 2007). Kolen and Brennan (2004) provide detailed descriptions of the linking methods used in those 12 studies.

The goal of these linking methods is to put scores from each state test on the same NAEP scale. While each method is unique, the methods used by the AIR the USDE both involve a three-step process. In this process, as described by Ho and Haertel (2006b), researchers first examine state test scores from the sample of students and schools that were administered the NAEP. In the second step, the researchers calculate the percentage of these students who are proficient or above on the state test. The final step is to find the NAEP cut score that sets the same percentage of students as proficient. This NAEP cut score then represents the state proficiency standard mapped onto the NAEP scale.

Ho and Haertel (2006) note that, "all else being equal, states that report greater percents proficient will have lower mapped standards" (pp. 2-3) and that, "higher scoring NAEP states will have higher mapped standards" (p. 3). The researchers concluded that,

"The mapping essentially penalizes state performance standards for reporting high percents of proficient students without commensurately high NAEP performance" (p. 3). Thus, mapped proficiency standards that are relatively low on the NAEP score scale may represent states that have manipulated their scoring standards (lowering the standard for proficiency or making the test easier) in order to inflate their test scores.

The AIR (McLaughlin et al, 2000) and USDE (Braun & Qian, 2007) methods both found large differences among the proficiency standards used by states. In employing their method to analyze 2005 data, Braun and Qian (2007) found that state proficiency standards varied widely. For grades 4 and 8 in reading and mathematics, the state proficiency standards spanned 60 to 80 score points on the 500-point NAEP scale. The researchers also found:

> a strong negative correlation between the proportions of students meeting the states' proficiency standards and the NAEP score equivalents to those standards, suggesting that the observed heterogeneity in states' reported percents proficient can be largely attributed to differences in the stringency of their standards (p. iii)

Thus, if these linking methods provide valid results, it appears as though states reporting higher percentages of proficient students are manipulating score standards in order to inflate their scores.

Stoneberg (2007) provides a guideline that suggests that *none* of the single-year comparisons described in this section should be used to estimate the impact of manipulations. Stoneberg noted that in 2002, an Ad Hoc committee from the National Assessment Governing Board (NAGB) recommended that comparisons between state test and NAEP results, "should not be conducted on a 'point-by-point' [single-year] basis" (p. 6) because of the potential impact of the differences between state and NAEP testing programs (to be discussed later). Because of these flaws, single-year methods are not ideal methods to use to compare state and NAEP results

Trend Comparisons of State and NAEP Results

As an alternative to single-year comparisons, an Ad Hoc Committee convened by the NAGB recommended that NAEP achievement levels be used as evidence to confirm *trends* in state test scores in 4th and 8th grade reading and mathematics (Ad Hoc Committee, 2002). The National Academy of the Sciences also recommended comparing trends in scores on state and NAEP tests, suggesting that comparisons should focus on changes in the percentages of students scoring proficient rather than focusing on results from a single year (Pellegrino, Jones, & Mitchell, 1998). As will be discussed, discrepancies between state and NAEP results from a single year can be influenced by differences in test content and the motivational levels of examinees. Some researchers believe these differences are not as problematic when comparing score trends. Linn, Baker, & Betebenner (2002) noted that, "Despite differences in the stakes attached to the results of state tests and measures such as NAEP in content coverage, it is relevant to ask the degree to which gains on a state test generalize to gains on other measures of achievement" (p. 6). Klein, Hamilton, McCaffrey, and Stecher (2000) also believed that trend comparisons addressed some of the problems with single-year comparisons, noting that "any reduction in student effort or performance that may stem from NAEP being a relatively low-stakes test should be fairly consistent over time and therefore not bias our measurement of score improvements across years" (p. 4). Linn (2000) further justified the trend comparison method, noting that, "Divergence of trends does not prove that NAEP is right and the state assessment is misleading, but it does raise important questions about the generalizabilty of gains reported on a state's own assessment, and hence about the validity of claims regarding student achievement" (p. 14).

Reports from The Thomas B. Fordham Foundation (2005), RAND (Klein, Hamilton, McCaffrey, & Stecher, 2000), and other researchers (Koretz & Barron, 1998; Linn, 2000; Linn, Baker, & Betebenner, 2002) have employed this method of trend comparison to estimate the impact of test score manipulations on state test results. In

comparing scores on the Kentucky Instructional Results Information System (KIRIS) to NAEP results, Koretz and Barron (1998) found that the large KIRIS gains reported for fourth grade reading and mathematics from 1992 to 1994 were more than four times larger than the gains in NAEP results over the same time period. The researchers concluded that the large KIRIS gains were due to teachers manipulating the teaching process to teach only the content of previous tests.

Klein, Hamilton, McCaffrey, and Stecher (2000), on behalf of RAND, and Linn, Baker, and Betebenner (2002) similarly compared trends from the Texas Assessment of Academic Skills (TAAS) to the NAEP. Both sets of researchers found that score gains reported from the TAAS were significantly larger than the gains reported from the NAEP. Figure 2.3 illustrates the findings. Figure 2.3a displays the score trends on the TAAS compared to trends on the NAEP in 8th grade mathematics from 1990 until 2001. The slopes of the test scores over time represents test score trends. Whether using the NAEP proficient or basic standards, Figure 2.3a shows that trends in TAAS pass rates outpace the NAEP trends. As a counterexample, Figure 2.3b shows a similar comparison between NAEP results and scores from the Maryland School Performance Assessment Program (MSPAP) in 8th grade mathematics. Maryland experienced trends similar to Texas on the NAEP over this time period. While TAAS results showed much greater growth than the NAEP, the trends from the MSPAP appear to support NAEP trends. Klein, Hamilton, McCaffrey, and Stecher (2000) concluded that the discrepancy in score trends in Texas could be attributed to:

> (1) students being coached to develop skills that are unique to the specific types of questions that are asked on the statewide exam (i.e., as distinct from what is generally meant by reading, math, or the other subjects tested); (2) narrowing the curriculum to improve scores on the state exam at the expense of other important skills and subjects that are not tested; (3) an increase in the prevalence of activities that substantially reduce the validity of the scores

In other words, discrepancies in score trends could be due to manipulations of the teaching philosophy or process or manipulations of the test administration.

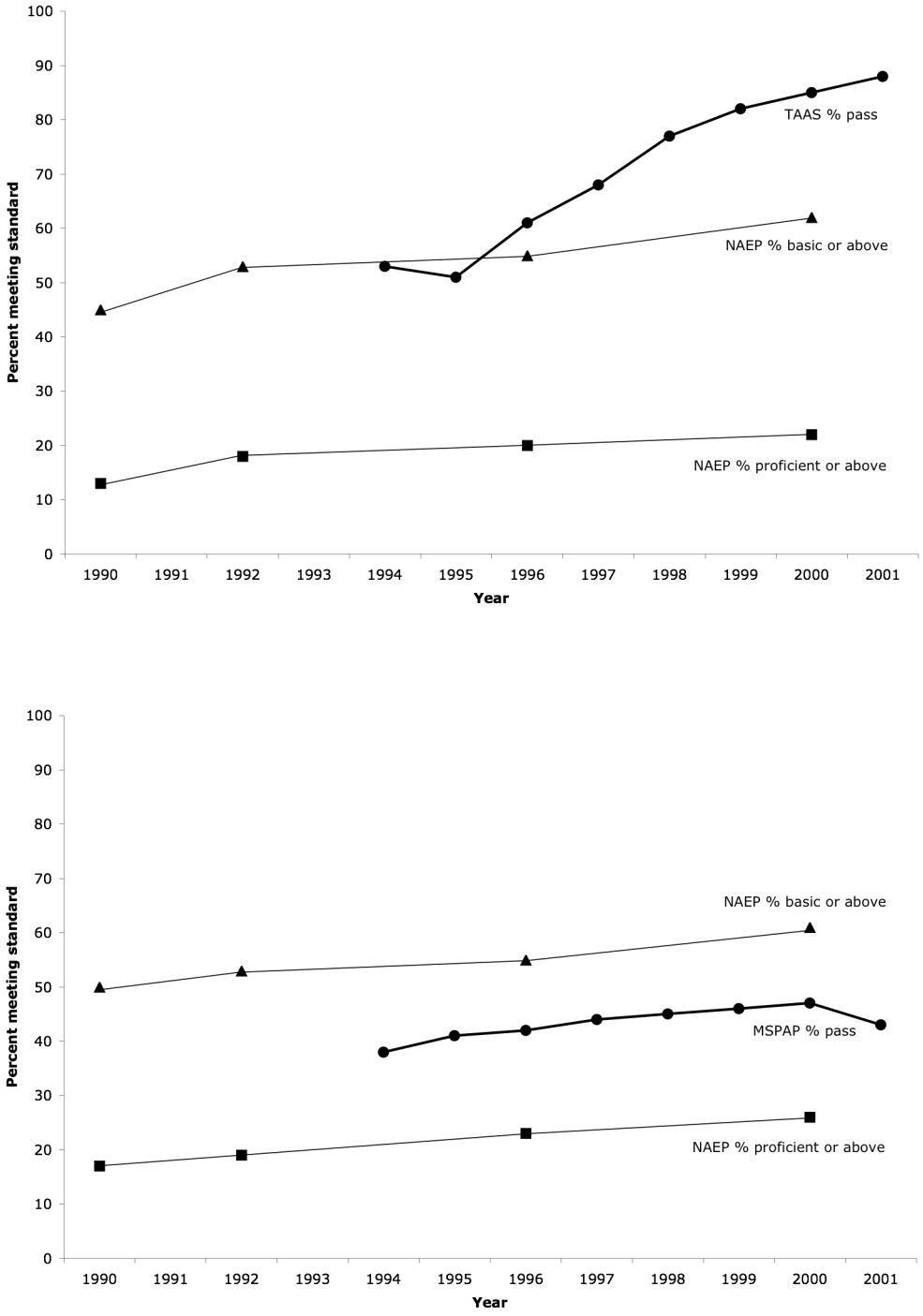


Figure 2.3 Score trends in 8th grade mathematics measured by the MSPAP, TAAS, and NAEP.

Source: Linn, Baker, & Betebenner (2002)

Jacob (2007) extended the research into discrepancies between score trends on the TAAS and NAEP. Jacob found that between 1996 and 2000, TAAS scores increased at a much higher rate than NAEP scores. He found, for example, that math performance increased by more than 0.5 standard deviations on the TAAS compared to a 0.1 standard deviation increase on the NAEP. Using item-level data, Jacob concluded that the discrepancy in score trends could not be explained by changes in the demographic composition of examinees or differences in test item formats. The fact that the NAEP is administered with a time limit whereas the TAAS is not timed also did not explain the differences in score trends. After conducting the analysis, Jacob expressed concern about the generalizability of student achievement gains under state accountability systems.

The Thomas B. Fordham Foundation (2005) conducted an analysis to compare trends on other state tests to the NAEP. Tables 2.12 and 2.13 illustrate a sample of this data. Table 2.12 shows that of the 29 states with reported score trends in 8th grade reading achievement from 2003-2005, 19 reported increases in the percent scoring proficient on the state tests. Of those 19 states, none experienced an increase in the percent proficient on the NAEP. In fact, 14 of those states experienced declines in proficiency on the NAEP. Table 2.13 shows that only three states experienced trends of the same direction in state test and NAEP results. Similar results in reading for grades 4 and 8 led the researchers to conclude that manipulations of the scoring standards inflated state test results. Fordham Foundation president Chester E. Finn, Jr. stated, "If states ease their standards, construct simple-minded tests, or set low passing scores, they can mislead their own citizens and educators into thinking that just about everyone is proficient" (p. 1).

Table 2.12 Discrepant trends in 8th grade reading achievement (2003-2005)

	Change in % **proficient** from 2003-2005 on…		Change in % **basic** from 2003-2005 on…	Change in % proficient on state test is greater than change in % ___ on NAEP	
State	State Test	NAEP	NAEP	Proficient	Basic
Alabama	11	0	-2	*	*
California	9	-1	-1	*	*
Idaho	9	0	0	*	*
Arizona	8	-2	-1	*	*
Delaware	8	-1	3	*	*
Tennessee	8	0	2	*	*
Maryland	6	-1	-2	*	*
Virginia	6	0	-1	*	*
Kentucky	5	-3	-3	*	*
Indiana	3	-5	-4	*	*
Iowa	3	-2	0	*	*
New York	3	-2	0	*	*
North Dakota	3	-1	2	*	*
Oregon	3	0	-1	*	*
Georgia	2	-1	-2	*	*
North Carolina	2	-2	-3	*	*
Oklahoma	2	-5	-2	*	*
Missouri	1	-3	-3	*	*
South Dakota	1	-4	0	*	*
Colorado	0	-4	-3	*	*
Mississippi	0	-4	-5	*	*
Wisconsin	0	-2	0		*
Wyoming	0	2	2		
Hawaii	-1	-4	-3	*	*
Maine	-1	0	2		
Connecticut	-2	-4	-3	*	*
Florida	-5	-2	-2		
Texas	-5	0	-2		
Montana	-6	0	0		
Median	**2**	**-2**	**-1**		

Source: Thomas B. Fordham Foundation (2005)

Table 2.13 Discrepant trends in 8th grade reading achievement (2003-2005)

		2003-2005 **State Test** Trend		
		Decline in % proficient	No change in % proficient	Growth in % proficient
2003-2005 **NAEP** Trend	Decline in % proficient	Hawaii Florida Connecticut	Colorado Wisconsin Mississippi	California Arizona Delaware Indiana Maryland Iowa Kentucky S Dakota Georgia Oklahoma Missouri N Carolina New York N Dakota
	No change in % proficient	Montana Maine Texas		Alabama Idaho Oregon Tennessee Virginia
	Growth in % proficient		Wyoming	

Unfortunately, these trends comparison methods are limited due to both substantive and technical issues. In explaining the technical issues, Ho (2007) describes "the act of comparing state and NAEP results as the act of comparing the height of two children on pogo sticks" (p. 2). When researchers measure trends in the percentage of students scoring above a cut-score, the magnitude and sign of those Percent Above Cut (PAC-based) trends depend on the selection of cut-score. As Ho explains:

> The interpretive problems with PAC-based statistics may be simply explained by their interaction with unimodal distributions. If a unimodal distribution of test scores shifts in the positive direction, the rate at which examinees cross a cut-score will not be constant. As the mode of the distribution approaches the cut-score, more and more examinees will cross in equal units of time. After the mode of the distribution passes the cut-score, fewer and fewer examinees will cross in equal units of time. If the cut-score were different under this model, the trend would be different. In this sense, PAC-based trends may be described as *pliable* under the choice of cut-score. (p. 4)

Ho goes on to demonstrate the pliability in PAC-based trends for state 4th grade reading results on the NAEP from 2003 to 2005 by calculating these PAC-based trends from the Basic, Proficient, and Advanced cut-scores on the NAEP. He finds, for example, that Arizona experienced a 1% gain in students scoring above the NAEP Advanced cut-score from 2003 to 2005. Using a different cut-score, Arizona experienced a 2% decline in students scoring above the NAEP *Basic* cut-score. Other states showed similar results in that the choice of cut-score changes the magnitude and sign of the trend in students scoring above that cut-score. Because conclusions from trend comparisons change depending on the selection of cut-score, these trend comparisons should be interpreted cautiously.

Scale-Invariant Trend Comparison Methods

To address the pliability of PAC-based trends, Ho (2007) introduced a scale-invariant trend statistic based on the Probability-Probability (PP) plot of score distributions from a test given at two times. The *V* statistic (Ho & Haertel, 2005) is

described as a scale-neutral effect size – a measure of the change in test scores from one time to the next that does not change depending on the selection of a cut-score. A detailed explanation of the *V* statistic will be provided in the next chapter.

Ho (2007) estimated the *V* statistic for 82 combinations of test results from 4th and 8th grade state and NAEP reading and mathematics tests from 2003 and 2005. The researcher found that the average trend in state test scores was significantly more positive than the average trend in NAEP scores, with 76% of the state trends being more positive than NAEP trends. After cautioning readers that these findings could be influenced by content differences, examinee motivation, examinee sampling, or other reasons, Ho concluded, "These results are consistent with the hypothesis that increased attention to state test content leads to improved performance on state tests but not on NAEP" (p. 13).

State and NAEP Trend Discrepancies: Plausible Rival Hypotheses

While researchers have concluded that manipulations may have caused discrepancies between state and NAEP results (Hall & Kennedy, 2006; Jacob, 2007; Kleine, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998; Lee, 2006; Peterson & Hess, 2005, 2006; Ravitch, 2005), it must be noted that the existence of discrepancies does not prove the existence or impact of test score manipulations. Jacob (2007) notes, "… there has been little research on reasons *why* student performance differs between NAEP and local assessments" (p. 11). This research is important because, as Koretz (1991) states, in order to conclude that a discrepancy between state and NAEP results, "reflects specific policies or practices, one needs to be able to reject with reasonable confidence other plausible explanations…" (p. 20).

Hill (1998), Ho and Haertel (2007), the Iowa Department of Education (2007), Jacob (2007), and Koretz (1999) all address plausible rival hypotheses that may explain

any discrepancies between state and NAEP results. Synthesizing this research, some of these plausible rival hypotheses include:

- Differences in content coverage or sequence or opportunity to learn
- Differences in item formats or administration mode (paper- or computer-based)
- Differences in test difficulty
- Differences in score standards or standard-setting procedures
- Differences in test administration procedures/environment or administration date
- Differences in accommodations allowed during testing
- Differences in examinee populations or subgroup definitions
- Differences in examinee motivation or effort

The first four plausible rival hypotheses address differences between state and NAEP tests and scoring procedures. If a state test differs from NAEP in content coverage or sequence, then it would be expected that students would score higher on the state test (due to educators focusing on state content standards). Likewise, differences in item formats, test administration mode, or test difficulty may have a significant impact on score discrepancies between state tests and NAEP. Also, as was discussed previously, score standards may also impact state-NAEP discrepancies.

The next two rival hypotheses for score discrepancies address differences in test administration procedures. If state test administration procedures significantly differ from NAEP procedures (in terms of testing time, use of accommodations, or use of materials such as calculators during testing), then discrepancies in results between the two tests would not be completely unexpected.

The final three possible explanations for score discrepancies deal with potential differences in the examinees being tested under state tests and NAEP. While NCLB requires at least 95% of students to be tested annually and NAEP sets its standard at 85% (Hill, 1998, p. 3), this means that up to 20% of examinees could have been excluded from at least one of the tests. Furthermore, NAEP participation guidelines very by state,

especially with regards to students with disabilities and English Language Learners. Clearly, these potential differences in the examinee pool could impact discrepancies between results from the two tests. Also, since state tests under NCLB are high-stakes and NAEP remains a relatively low-stakes test, differences in examinee motivation or effort could have an impact on discrepancies.

These plausible rival hypotheses are not exhaustive, but they do provide a reminder that discrepancies between state and NAEP score trends do not automatically mean that educators have manipulated test scores. In order to have confidence in a causal relationship, strong assumptions must be made that the discrepancies are not due to the above plausible rival hypotheses. While several studies have concluded that differences in test content (Wei, Shen, Lukoff, Ho, & Haertel, 2006), examinee motivation (Klein, Hamilton, McCaffrey, & Stecher, 2000; Linn, Baker, & Betebenner, 2002), examinee demographics, test item formats, and test administration time limits (Jacob, 2007) cannot explain the discrepancies between state test and NAEP results, the existence of these differences should at least temper expectations about the comparability of state test and NAEP score trends.

Another assumption implicitly made in comparing state and NAEP score trends is that results from NAEP are somehow the "gold standard." While NAEP may not be the gold standard, it may be the only available standard with which to compare the performance of all states in reading and mathematics achievement. While NAEP scores have been more robust against educator manipulations, Hill (1998) notes, "As more and more states see the need for increased NAEP scores, practices will evolve that will virtually ensure gains on NAEP" (p. 10). Thus, Hill suggests that if NAEP results are used to validate state test results, NAEP results will become high-stakes and NAEP will become subject to the same manipulations as state tests.

While discrepancies between state and NAEP score trends cannot be attributed to educator manipulations, the relationship between the quality of a state's test security

policy and the magnitude of state-NAEP discrepancies is still interesting. If an inverse relationship is found between policy quality and discrepancy magnitude, future research could be targeted to determine if those discrepancies could possibly have been caused by manipulations.

Summary

This literature review has shown that educators do manipulate test scores by manipulating the teaching philosophy or process, manipulating the examinee pool, manipulating the test administration, or manipulating score reports or standards. While the exact prevalence of each form of manipulation is unknown, the evidence suggests that reported incidents of manipulations are widespread and increasing.

This literature review has also shown that educators might manipulate test scores because of a lack of effective test security policies at the state level. Using research into student cheating and honor codes along with test security survey results, this literature review suggests that states can foil test score manipulations by formalizing testing beliefs and practices, overseeing test activities, informing educators about what behaviors are appropriate and inappropriate, and limiting opportunities for educators to manipulate test scores.

Finally, while cautioning against causal interpretations and providing a list of some plausible rival hypotheses, this literature review makes the case that the relationship between the quality of a state's test security policy and the magnitude of discrepancies in trends between the state test and NAEP is of interest.

CHAPTER 3: METHODOLOGY

The purpose of this study is to determine if a relationship exists between the existence/quality of state test security policies and discrepancies between state test and NAEP score trends. Specifically, this study attempts to address the following research questions:

1. What kinds of manipulations do educators use to increase test scores? Why do educators manipulate test scores? What is the estimated prevalence of each type of manipulation?
2. What test security policies and practices do states implement in an attempt to deter educators from manipulating test scores? What is the quality of each state's test security policy?
3. What is the relationship between the quality of a state's test security policy and any discrepancies between score trends on state and NAEP tests? Which aspects of a state's test security policy seem to have the strongest relationship with score trend discrepancies? What are some potential explanations for the discrepancies between state test and NAEP score trends?

To address the first set of research questions, studies and news reports were synthesized to develop a taxonomy of test score manipulations. While the prevalence of each manipulation method could only be roughly estimated, the evidence suggests that incidents of manipulations are widespread and growing.

To address the second set of research questions, a FOIL framework was developed to evaluate the existence and quality of four aspects of test security policies. The quality of test security policies serves as the independent variable in this study.

To address the third set of research questions, a scale-invariant framework is used to compare trends between state tests and NAEP scores. Test score discrepancies (the dependent variable) and test security policies are then compared among states to

determine if states with higher quality test security policies experience smaller score trend discrepancies than states with lower-quality test security policies. When possible, comparisons are also made within states that recently adopted new test security policies or significantly modified existing policies to determine the relationship of these policies with score trend discrepancies.

Independent Variable: Test Security Policy Quality

The independent variable in this study is the quality of test security policies implemented by states to deter educators from manipulating test scores. First, information regarding state test security policies was collected and organized. Then, the quality of the test security policy content and implementation was evaluated using the FOIL framework developed in the previous chapter (displayed in Table 2.10). Each state's adopted policy was evaluated holistically and with regards to each of the four aspects of the FOIL framework.

In order to collect information about each state's test security policies, the term *policy* must be defined. According to the Stanford Policy Repository (2007), a policy is:

> a statement of *principles* and/or *values* that mandate or *constrain the performance of activities used in achieving institutional goals*. A policy is general in nature, has broad application and helps to *ensure compliance with: applicable laws and regulations*; contract requirements; and *delegation of authority*.... Policies promote operational efficiencies and *reduce institutional risk*. ... Directives, processes, procedures, work instructions, and the like flow from policies... (emphasis added).

Since the institutional goal of a state's public education system, according to the mandates of NCLB, is to increase student achievement as measured by test scores, a state test security policy is a written plan of action to guide educators' decisions and actions in testing. The test security policy guides these decisions and actions to ensure they comply with the state's values, principles, laws, and regulations.

A state test security policy must be contrasted with test administration manuals and the national testing codes and standards outlined in Appendix C. As explained in the

previous chapter, evidence indicates that administration manuals and national standards are ineffective in deterring educators from manipulating test scores. This ineffectiveness may be due, in part, to the overabundance of these materials (which sometimes offer conflicting guidance) and/or to the fact that these materials are often perceived as suggestions rather than mandates.

State test security policies differ from test administration manuals or national standards in that policies are mandates. For this study, a test security policy must be connected with state rules, regulations, laws, or sanctions to clearly demonstrate that the policy is a mandate. Also, for this study, a state test security policy must be issued or endorsed by a state Department of Education (DOE), Board of Education (BOE), or State Legislature (SL). This will differentiate state security policies from guidelines developed by schools, school districts, or local education agencies.

Data Collection and Verification

State test security policy information was collected from publicly available information published on state DOE, BOE, and SL websites. These sites were navigated to find information regarding state assessment programs. Also, these sites were searched for key phrases such as *test security, test policy*, *test guidelines,* and *ethics codes*. Any information regarding the state's testing principles, requirements, laws, and regulations was collected. The data were then entered into a standardized form (see Figure 3.1). If data from any state were missing, incomplete, or contradictory, state DOE officials were contacted for verification.

Sampling

While it was not necessary to collect test security policy information from states excluded from the analysis (see Table 3.6), policy information was collected from 493 documents from all 50 states. When available, information about changes to policies between the years 2003-2007 were collected for each state.

State test security policy evaluation form

State name: _____ Test Name: _____

Year policy adopted/modified: _____ Item format: _____

Rate each of the following according to the scale: 0 = Missing 1 = Meets 2 = Exceeds

Formalize beliefs of state educators regarding the role of testing and testing practices

• Prominence / Availability of information

_____ The state has a separate test security office or budget

_____ The state has a separate web page for information about test security

_____ The policy is mentioned in test administration manuals

_____ Percentage of test administration manual pages dedicated to test security = _____%

_____ Number of clicks to navigate from the front page to test security information = _____

• Content

_____ Teachers provided input into content (evidence of a committee or documentation of development)

_____ Clarity of test security policy information

_____ Availability of FAQs regarding test security

_____ The policy requires educator signatures to indicate understanding

_____ Amount of information available (number of documents = _____)

• Implementation

_____ Identifies an individual or group in charge of security: ___ individual or ___ group

_____ Identifies individuals responsible for security at both state and district levels

_____ Evidence of a dissemination plan being followed

_____ Policy content is updated regularly

_____ Policy provides for barcodes (or another system) to automate test score identification

_____ Availability of forms and checklists for districts/schools to use to aid in test security practices

• Requirements and sanctions

_____ Mandatory reporting requirements (for suspected incidents of manipulation)

_____ Provides standard forms (or online reporting) for suspected incidents

_____ Explains the protections for individuals who report suspected incidents

_____ Due process is explained (procedures for investigating suspected incidents)

_____ Sanctions for confirmed cases of manipulation are outlined

_____ Sanctions include suspension and dismissal of confirmed manipulators

• Other

_____ Explains the importance of test security: ___ positive message or ___ negative message

Figure 3.1 Evaluation form for state test security policies.

Figure 3.1 Continued

_____ Laws, regulations, rules are mentioned (to indicate that security is a mandate)

_N/A_ Test security policies are to be developed at: ___ state-level or ___ district-level

Oversee test preparation, administration, and scoring activities

• Test Security Audits

_____ Implementation of test security policy is audited regularly (evidence of improvements)

• Test administration oversight

_____ The policy provides for independent monitoring of test administration

_____ Teachers are not to administer the test to their own students

• Statistical Analyses

_____ The policy provides for statistical analyses of answer sheets to detect possible manipulations

_____ The policy provides for erasure analysis

_____ The policy provides for aberrant response analysis

_____ The policy provides for analysis of score fluctuations

_____ Evidence of security analysis reports

• Score Reports

_____ The policy outlines procedures to follow before making any changes to test scores

Inform educators about why some behaviors and activities are unacceptable

• Principles & Rules

_____ The policy explains copyright laws and penalties for violating copyright

_____ The policy explains the importance of validity & generalizing from test scores

_____ The policy explains that all students must be tested (as required by NCLB)

_____ The policy explains the importance of standardized test administration

_____ The policy refers to an honor code or code of ethics

_____ The policy describes the uses of test scores

• Examples of appropriate and inappropriate behaviors

_____ Examples of test preparation activities: _____ appropriate and _____ inappropriate

_____ Examples of test administration activities: _____ appropriate and _____ inappropriate

_____ Examples of accommodations: _____ appropriate and _____ inappropriate

_____ Examples of school/class activities on test day: _____ appropriate and _____ inappropriate

_____ Examples of uses/interpretation of scores: _____ appropriate and _____ inappropriate

• General Guidance

_____ The policy limits the amount of time spent on test preparation activities

_____ The policy specifically states that educators cannot change student answers

Figure 3.1 Continued

_____ The policy specifically states that educators cannot give students hints or answers

_____ The policy specifically states that educators cannot read certain sections aloud to students

_____ The policy lists what materials teachers can or cannot provide to students during testing

_____ The policy outlines procedures for retesting students or sanitizing answer sheets

• Training

_____ The policy provides for regular training of district-level testing coordinators

_____ The policy provides for regular training of school-level testing coordinators

_____ The policy provides for regular training of all test proctors

_____ The quality of training materials available online

_____ Regularity / amount of training

Limit opportunities for educators to manipulate test scores

• Materials security

_____ The policy specifies who has access to test materials (and at what times)

_____ The policy specifies the amount of time materials are available

_____ The policy provides for a tracking system of test materials

_____ The policy requires test materials to remain sealed until testing

• Test Forms

_____ The policy requires new test forms (different test items) to be used annually

_____ The policy requires multiple test forms (which may be reused)

Overall categorization of test security policy:

A)	___ Clear and accessible	vs.	___ Ambiguous or difficult-to-find
B)	___ State-level mandate	vs.	___ District- or school-level responsibility
C)	___ Punitive or law-focused	vs.	___ Instructive/informative
D)	___ Independent monitoring	vs.	___ No independent monitoring
E)	___ Investigative	vs.	___ Preventative
F)	___ Example-based	vs.	___ Not many examples
G)	___ Positive message	vs.	___ Negative message

Additional Information:

Analysis

Once policy information had been collected, the quality of each policy was evaluated according to the FOIL framework. Figure 3.1 shows the form that was used to evaluate the quality of more than 60 features of each state's policy according to a 3-point scale. A score of 0 was given to any state policy that was missing the feature; a score of 1 indicates the state policy contained the feature; a score of 2 indicates the state policy was exceptional in that feature. This scoring method produced a 122-point composite scale for test security policy quality.

The evaluation form contains additional information about each state's test security policy, including the year in which the policy was developed or modified, and the format of the items on the state test. The year of policy development/modification was used, when possible, for longitudinal analyses of policy effectiveness. Item format information was used to test the hypothesis that states administering tests similar in format to the NAEP experience smaller trend discrepancies between the two tests.

The evaluation form was also used to collect the following information:

- Percentage of pages in state test administration manuals that are focused on test security information
- The number of test security policy documents available online
- Whether the policy identifies an individual or a group of individuals (committee) responsible for test security
- Whether the policy explains the importance of test security via a positive or negative (punitive) message
- Whether the policy is developed at the state- or district-level
- The number of examples of appropriate and inappropriate testing activities

The results for each feature along with this additional information were used to dichotomize state policies in 7 ways:

- Clear and accessible (specific information is readily available) vs. ambiguous or difficult to find
- State-level mandates vs. district- or school-level responsibility (states that require districts to develop test security policies)
- Punitive (focused on sanctions) vs. instructive (focused on informing educators on the importance of test security)
- Policies that require independent monitoring of test administration vs. policies that allow teachers to administer tests to their own students
- Investigative (focused on reporting and investigating potential manipulation incidents) vs. preventative (focused on preventing manipulations)
- Example-based (provides many examples of appropriate and inappropriate behaviors) vs. general (provides general information without specific examples)
- Positive message (provides examples of positive testing behaviors) vs. negative message (provides examples of negative testing behaviors).

A short narrative provides additional information collected from test security policy documents.

Once all state policies had been examined, the data were summarized. One way to summarize this information was to report the number of states receiving scores of 0, 1, and 2 for each feature. These scores were also summed to produce composite scores for each component (formalize, oversee, inform, and limit) and subcomponent (prominence, content, implementation, etc.) displayed in Figure 3.1. The distributions of these composite scores were examined and medians were computed for each section and subsection. These composite scores were also computed for each state. Figure 3.2 displays the composite scores that were calculated.

State test security policy evaluation form

State name: _____ Test Name: _____

Year policy adopted/modified: _____ Item format: _____

Composite Scores

_____ **Formalize** beliefs of state educators regarding the role of testing and testing practices
- _____ Prominence / Availability of information
- _____ Content
- _____ Implementation
- _____ Requirements and sanctions
- _____ Other

_____ **Oversee** test preparation, administration, and scoring activities
- _____ Test Security Audits
- _____ Test administration oversight
- _____ Statistical Analyses
- _____ Score Reports

_____ **Inform** educators about why some behaviors and activities are unacceptable
- _____ Principles & Rules
- _____ Examples of appropriate and inappropriate behaviors
- _____ General Guidance
- _____ Training

_____ **Limit** opportunities for educators to manipulate test scores
- _____ Materials security
- _____ Test Forms

Overall categorization of test security policy: Check one for each row

A)	___ Clear and accessible	vs.	___ Ambiguous or difficult-to-find
B)	___ State-level mandate	vs.	___ District- or school-level responsibility
C)	___ Punitive or law-focused	vs.	___ Instructive/informative
D)	___ Independent monitoring	vs.	___ No independent monitoring
E)	___ Investigative	vs.	___ Preventative
F)	___ Example-based	vs.	___ Not many examples
G)	___ Positive message	vs.	___ Negative message

Figure 3.2 Example of composite scores calculated for each state.

Technical Quality of Policy Evaluation Data

As stated in the previous chapter, the components of the test security policy evaluation are based on suggestions from test developers and publishers (Harcourt

Assessment, 2007; Iowa Testing Programs, 2005; Riverside Publishing, 2006); research into test preparation activities (Crocker, 2003, 2006; Gay, 1990; Lai & Waltman, 2007; Moore, 1994; Popham, 1991), research into statistical detection of aberrant responders (Bellezza & Bellezza, 1989; Jacob & Levitt, 2003, 2004; Qualls, 2001; Sorensen, 2006; Wesolowsky, 1990), and research into student and teacher cheating (Cizek, 1999, 2003; Cullen & Reback, 2006; Figlio & Getzler, 2002; Jacob, 2007; McCabe & Trevino, 2002). Through discussions with a district assessment director from an Iowa public school district, one university assessment coordinator, and two (faculty) experts in educational measurement, the evaluation form was further refined.

Assumptions and Limits

The primary assumption with regards to the independent variables in this study is that the quality of a state's test security policy can be inferred from the quality of the materials and information made publicly available. This might be problematic, especially when trying to evaluate the quality of policy implementation based on available information. Just because a state has an exceptional policy (and materials describing the policy) does not mean that the policy was implemented well. An in-depth case study of a single state (or small group of states) would need to be conducted to evaluate the quality of implementation.

The main limitation with regards to the independent variables is that some information may not be available. It may be that some states lack information about certain aspects of their test security policies. With 493 policy documents collected from all 50 states, this did not become a problem. Another limitation is the lack of an interval scale on the data collection form. Since each aspect of the security policies are rated on a 0-1-2 scale, the resulting composite scale cannot safely be assumed to have interval scale properties. A third limitation is the subjective nature of some of the ratings assigned to state policies. While many of the 60-plus features can be objectively scored, features

such as the clarity of a policy or the quality of materials to train test proctors were subjectively rated. To address this, the rubric used to rate states in these subjective features was more clearly specified once the data had been collected.

Dependent Variable: Scale-Invariant State Test and NAEP Score Trend Discrepancies

The discrepancy in trends between state test and NAEP scores is used as the dependent variable in this study. Ideally, traditional effect sizes (mean differences divided by a pooled standard deviation) for both state and NAEP score trends would be compared to estimate the discrepancies. Unfortunately, as Jacob (2007) also discovered, "many states do not even publish state level averages of the underlying raw or scaled score, but rather report student performance in terms of the percent meeting various proficiency levels" (p. 14). Because of this limitation in the data (and because of the limitations in simply comparing proficiency rates between the two tests), a scale-invariant method (to be described later in this chapter) was used to estimate the discrepancies in effect-sizes between state and NAEP score trends.

First, results from state and NAEP testing in 2003, 2005, and 2007 were collected. Then, scale-invariant effect sizes, V_{state} and V_{NAEP} (to be described later in this chapter), were used to measure trends in state and NAEP results from 2003-2005, 2005-2007, and 2003-2007. The simple difference between V_{state} and V_{NAEP} define the state-NAEP score trend discrepancies. Table 3.1 displays the data that was collected and estimated for each state. With 50 states, 2 subjects (reading and mathematics), 2 grade levels, and 3 trends, a maximum of 600 scale-invariant discrepancy estimates could have been collected. The final sample in this study consists of 215 discrepancy estimates (36% of the maximum possible) from 32 states.

Table 3.1 Scale-Invariant Trend Discrepancy Data

			2003-05 score trends	2005-07 score trends	2003-07 score trends
State	Reading	4th Grade	$V_{state} - V_{NAEP}$	$V_{state} - V_{NAEP}$	$V_{state} - V_{NAEP}$
		8th Grade	$V_{state} - V_{NAEP}$	$V_{state} - V_{NAEP}$	$V_{state} - V_{NAEP}$
	Mathematics	4th Grade	$V_{state} - V_{NAEP}$	$V_{state} - V_{NAEP}$	$V_{state} - V_{NAEP}$
		8th Grade	$V_{state} - V_{NAEP}$	$V_{state} - V_{NAEP}$	$V_{state} - V_{NAEP}$

Data Collection and Verification

Online resources were searched to collect state and NAEP test scores. To collect results from state tests, the website of the Council of Chief State School Officers (CCSSO) was first explored to collect general information about each state's testing program. For each state, the following information for grades 4 and 8 in reading and mathematics for the years 2003, 2005, and 2007 was collected: the name of the test used for accountability requirements of NCLB, the test format (norm-referenced, criterion-referenced, or other), test administration dates (fall or spring), and the number of cut-scores and performance-levels reported. The years 2003, 2005, and 2007 were chosen because they correspond with NAEP administration years. After entering this information into a text file, any changes made to a state's testing program in 2005 or 2007 were flagged.

Figure 3.3 displays an example of the data collected from the CCSSO website. From this figure, one can see that Alabama changed from a norm-referenced to a criterion-referenced test between 2003 and 2005 and that the state testing program classifies students into four performance levels (two of which represent proficient levels of performance). Because Alabama changed tests, this state was flagged to indicate that no valid trend comparisons exist from 2003-2005 or from 2005-2007.

After collecting and entering information from the CCSSO website for all state testing programs, this information was verified by examining the websites of each state's

Department of Education (State DOE). From these State DOE websites (and web searches) additional data about each state's testing program was entered. The general test information collected is displayed in Table 3.2. Once again, any changes made to a state's testing program were documented so that inappropriate trends (trends in scores on different tests or from tests administered at different times in the year) were not computed.

State: Alabama

Overview: CCSSO Database: http://accountability.ccsso.org/state_profiles.asp

Test Information

	2003	2005	2007
4th Grade Reading	Stanford 10	ART	ART
4th Grade Math	Stanford 10	AMT	AMT
8th Grade Reading	Stanford 10	ART	ART
8th Grade Math	Stanford 10	AMT	AMT

Notes: Stanford 10 was a norm-referenced test administered in April of 2003
ART = Alabama Reading Test
AMT = Alabama Mathematics Test
ART and AMT were administered in April of 2005 and 2007

The state uses 3 cut-scores to classify students into the following categories:

Level 1	Level II	Level III	Level IV
(Below proficient)		(Proficient)	

Figure 3.3 Example of data collected from CCSSO.org website.

Once general information about state testing programs had been collected and entered, state test results were collected. At a minimum, the percentage of students scoring in each performance level in reading and mathematics in grades 4 and 8 in 2003, 2005, and 2007 were collected. With this data, the percentage of students scoring at or

above proficient in each subject and grade level in each year were calculated. If this minimum amount of data was not readily available online, calls or emails were made to state DOE officials to request the information. When available online, additional information was collected for each grade and subject, such as the number of students tested each year and the scale score means and standard deviations (to calculate traditional effect sizes). Table 3.3 shows an example of the test score data that was collected for each state.

Table 3.2 General test information collected from each state DOE website

Information	Notes
Test name and format	When both are available, results from *reading* tests are preferred to *English* or *Language Arts* tests. When available, item types (selected-response, constructed-response, or mixed) are recorded. Also, when possible, results from alternate assessments were excluded from analysis.
Test Administration Date	Fall or Spring
Number of Cut-Scores	Cut-scores are used to classify students into performance levels
Proficiency Levels	Which performance levels that represent proficiency
Changes	Any changes in test name, format, administration date, cut-scores, or performance levels are flagged.

The table shows, once again, that Alabama changed tests between 2003 and 2005. This change means that the 2003 test scores were used to measure trends from 2003-2005 or 2005-2007. Because of this, the 2003 test score data was collected. If a state changed tests after 2003 and again after 2005, then only the 2007 data was collected as no appropriate trends could be measured.

The table also shows, once again, that the Alabama testing program used 3 cut-scores to place students into one of 4 performance levels. The *Decimals* variable indicates with what precision the state reported its test results. In this example, Alabama

reported the percentage of students scoring in each performance level with two decimals of precision. The table shows that while the number of students tested in 2005 was not found online, a total of 56,083 4th grade students were administered the ART in 2007. The table also shows that the mean scale scores and standard deviations were not readily available online, but that the percentages of students scoring in each performance level were available. The percentage of students scoring at or above proficient was calculated by adding the percentage of students scoring in the third and fourth performance levels. Finally, the table also shows that scores were not adjusted in any way (from the reports available online) and that trends can only be made from the 2005 and 2007 data.

Table 3.3 Example of test score data to be collected

		Alabama		
		2003	2005	2007
4th Grade Reading	Test	Stanford 10	ART	ART
	Format	NRT	CRT	CRT
	Items	Constructed	Mixed	Mixed
	Administration	Spring	Spring	Spring
	Cut-Scores	3	3	3
	Decimals	--	2	2
	# Tested	--	N/A	56,083
	Avg. Score	--	N/A	N/A
	Std. Deviation	--	N/A	N/A
	% in PL 1	--	0.31%	0.50%
	% in PL 2	--	16.35%	14.49%
	% in PL 3	--	33.18%	31.86%
	% in PL 4	--	50.16%	53.15%
	% Proficient	--	83.34%	85.0%
	Adjusted?	--	No	No
	Cat Shift?	1	2	2

Notes: N/A = information not available online

PL = Performance level

Adjusted = Were scores adjusted in any way from what's reported online?

Cat Shift = Equal values represent years in which trends can be compared

To aid in data entry, a spreadsheet was developed and state-specific data collection issues were discussed with Educational Measurement and Statistics faculty. When available, the data entered into the spreadsheet were verified by other test score reports available online. The percentages of students scoring in each performance level were also verified by checking to see if the percentages sum to 100%. Some states had percentages that summed to either 99% or 101% due to rounding in the score reports. For these states, the percentages of students scoring within each performance level were divided by the sum of the percentages (scaled-up) to ensure all states had a sum of 100%.

Any changes in the testing program or problems in data entry were flagged for further investigation. As stated earlier, any changes in the test, test administration date, or cut-scores were flagged so that inappropriate trends would not be computed. When available, the numbers of students tested each year (for subgroups defined by race, disability status, socio-economic status, and English proficiency) were compared to check for any major changes in the testing population from year-to-year. The percentages of students scoring within each performance level and average scale scores were also visually examined to check for unusually large fluctuations.

For the flagged states, other online resources were searched and DOE officials were contacted to determine the explanation for the unusual data. These flagged states were also discussed with faculty during regular data collection meetings. From these meetings, decisions regarding which state test results to include or exclude from the analysis were made.

Sampling

Ideally, the data set would have included the percentage of students scoring within each performance level in grades 4 and 8 in reading and mathematics during the 2003, 2005, and 2007 administrations of the state tests. Due to incomplete or missing

data, or due to idiosyncrasies in some states' testing programs, cut scores, or score reports, some state data had to be excluded from analysis. The following rules were used to determine if data should be included in the analysis:

- All pairs of results (the 2003-2005, 2005-2007, and 2003-2007 pairs) were included if:
  - The state tested grades 4 and 8 in reading and mathematics both years.
  - The same test, or parallel forms, was administered both years.
  - Cut-scores and performance levels were not changed in either year.
  - The state administered the test during the same season (fall or spring) each year.
  - The population of students tested each year remained relatively stable (no wild fluctuations in the number of students within subgroups tested each year).
  - The state uses at least 3 cut-scores to place students into at least 4 non-overlapping performance levels.

Pairs of results were excluded from the analysis if the state changed the test or cut-scores (without equating the new scale to the old). Pairs were also excluded when the test administration date changed (from fall to spring or spring to fall) or when the data indicated that the tested student population had changed in some significant way. These pairs of results were excluded, because the test results may have different meanings in each year and, therefore, no appropriate trend could be computed. The last criteria requiring states to have at least 3 cut-scores and 4 non-overlapping performance levels was used only due to the requirements of the nonparametric estimation procedure that was used to analyze the data (explained later).

NAEP Data Collection

With the state test results collected, NAEP results were then collected from the official NAEP results website, *The Nation's Report Card* (2007ab). The *Trends in Achievement Levels by States* reports display the percentage of students scoring in each of the four NAEP performance levels: below basic, basic, proficient, and advanced. Through these reports, the percent of students scoring in each performance level were collected for each state in reading and mathematics in 2003, 2005, and 2007. The sums of the percentages were calculated as a check of the accuracy of data entry.

Analysis

Once state test and NAEP data had been collected, discrepancies in score trends for the two types of tests were estimated. As previously discussed, inadequacies in many state test score reports did not allow traditional effect sizes to be estimated. Also, as discussed in the previous chapter, many common methods to calculate discrepancies between state test and NAEP results have technical limitations. Specifically, single-year and trend comparisons of the percentage of students scoring at or above proficient are troublesome because of their pliability under different choices of cut-scores (to be illustrated later). In this study, a scale-invariant framework developed by Ho and Haertel (2006) was used with state test and NAEP results to estimate the discrepancy in score trends on the two tests.

Technical Limitations of Comparing Changes in Percentages of Proficient Students (PPS)

As has been mentioned previously, trend comparisons based on changes in the percentage of students scoring above a cut-score are known to be dependent on the choice of cut-score (Holland, 2002; Ho & Haertel, 2005; Koretz & Hamilton, 2006), which makes them of limited usefulness in comparing state-NAEP score trends. To illustrate this, Figure 3.4 illustrates score distributions from two simulated administrations of the

same test. The data were simulated so that from Time 1 to Time 2, the mean score increased from 550 to 600 and the standard deviation decreased from 150 to 100 from the first to the second administration. The data were simulated this way not only to provide a clear example, but also because the goal of NCLB is to both increase achievement for all students (increasing the overall mean) and decrease gaps in student achievement (perhaps decreasing the standard deviation).

Suppose the simulated data in Figure 3.4 come from a test with a cut-score of 500. The figure shows that 63% of students at Time 1 and 84% of students at Time 2 scored above this cut-score of 500. If this cut-score were defined as the proficiency standard, the state producing these results would be lauded for increasing proficiency by 21%. If, instead, a cut-score of 700 defined proficiency, the figure shows that the state would be viewed as ineffective in increasing achievement (26% of students scored above this cut-score at both Time 1 and Time 2). Using a cut-score of 800 (possibly reflecting higher expectations), a comparison of the percentage of proficient students (PPS) would lead to a conclusion that score trends were actually negative (proficiency dropped from 5% to 2%). Thus, this figure illustrates that the choice of cut-score can impact the conclusions drawn from PPS-based trend statistics.

Holland (2002) recommends using cumulative distribution functions to display the gap between two test score distributions. As explained by Wilk and Gnanadesikan (1968) CDFs provide a graphical display of a distribution's location, spread, and shape; and CDFs lend themselves to smoothing and interpolation. For a test administered at Time 1 and Time 2, a CDF provides a visual display of both:

$$F_1(x) = \%\text{ of students scoring at or below cut-score } x \text{ at Time 1} \quad (1)$$

and

$$F_2(x) = \%\text{ of students scoring at or below cut-score } x \text{ at Time 2}. \quad (2)$$

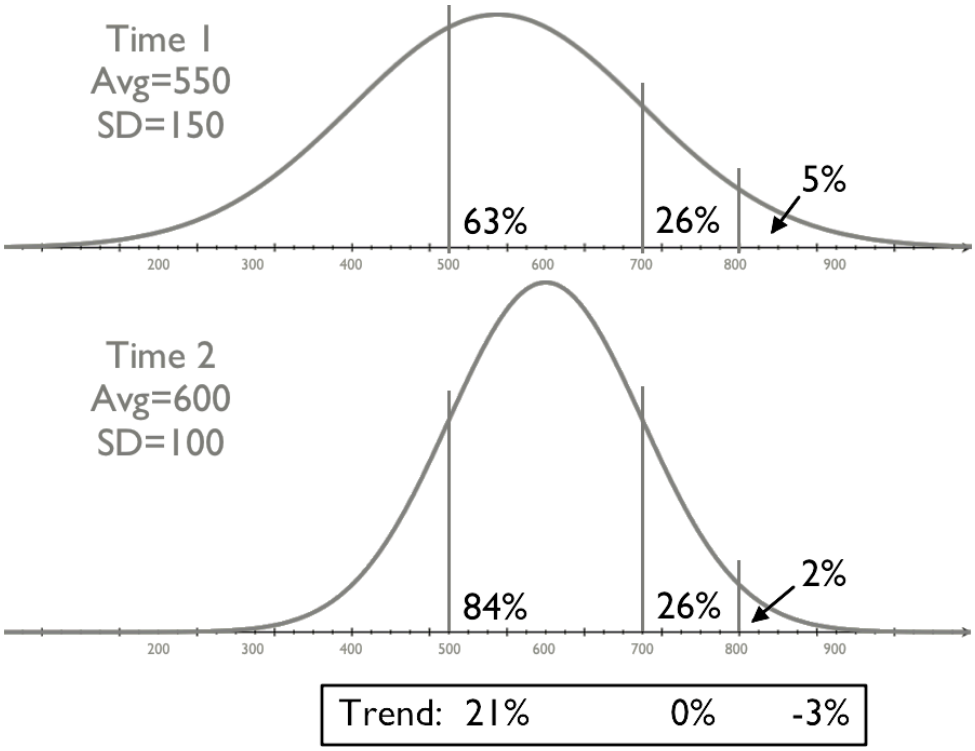


Figure 3.4 Simulated test score distributions to illustrate pliability of PPS-based trends.

Figure 3.5 displays CDFs for the same simulated distributions described earlier. The gap between the CDFs displays the trend in scores from Time 1 to Time 2. As Holland (2002) explains, this gap could be measured in several ways, most obviously by measuring the gap either horizontally or vertically. In his article, Holland recommended measuring the gaps between CDFs horizontally to represent the difference in percentiles from each distribution. The figure shows, for example, that the 50th percentile at Time 1 is equal to a scale score of 550. The 50th percentile at Time 2 is equal to a scale score of 600. Therefore, this horizontal gap displays a general positive trend in scores from Time 1 to Time 2 (at the 50th percentile).

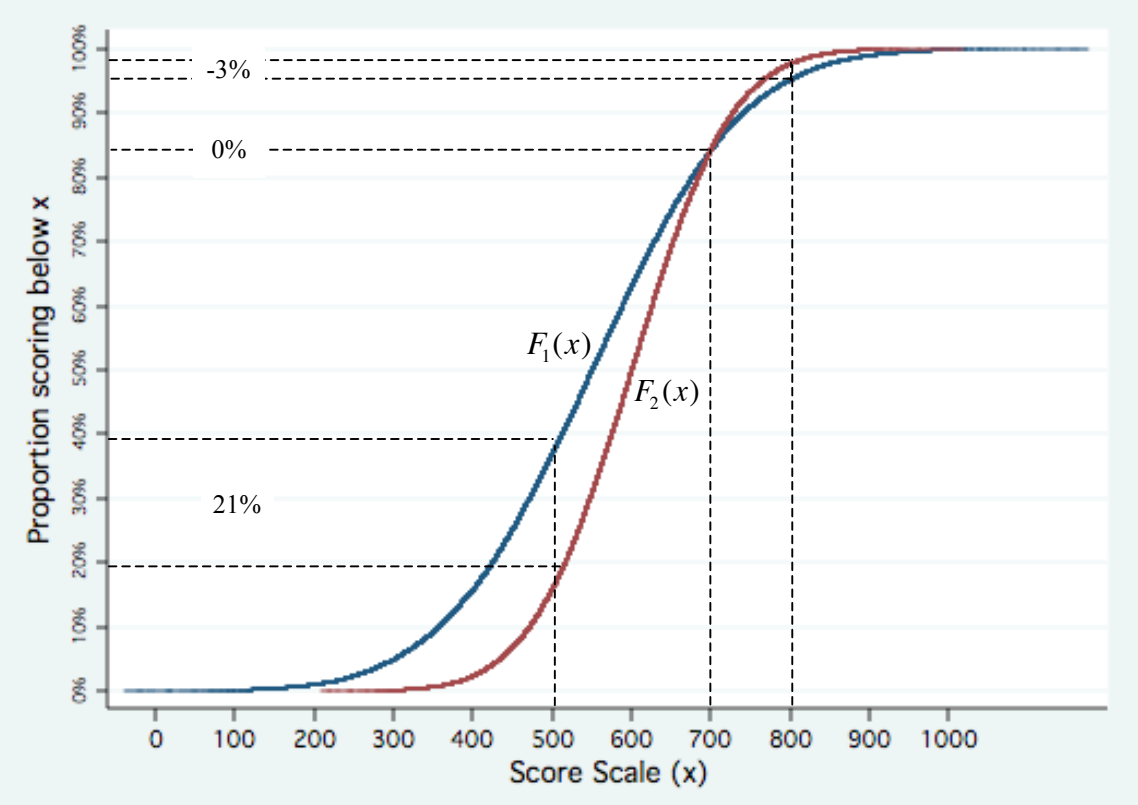


Figure 3.5 CDFs from a test administered at Time 1 and Time 2 with cut-scores of 500, 700, and 800

In the same article, Holland (2002) defended the use of vertical gap measurements in standards-based testing situations by explaining:

> The use of cut-scores is common among those interested in "standards-based" assessments. What could be more natural than to measure just how many in a group of examinees meet or exceed the cut-off for a standard with a name such as "Master" or "Proficient"? When educational reform is coupled with standards-based assessment, reformers are naturally led to ask if the percents of certain groups of students meeting or achieving a standard are increasing, possibly as a result of the reforms. From this point of view, the fact that score distributions often change over time by slow but steady shifts upwards is of little interest. The ultimate goal is to implement reforms in such a way as to get as many students as possible to meet or exceed the standards (p. 16).

Since NCLB is focused on standards-based assessment, the vertical gaps between CDFs are of interest in this study. The vertical gaps between the CDFs in Figure 3.5 once again show how the choice of cut-score impacts conclusions drawn from measuring the change in percentage of students scoring above a cut-score (PPS-based trends).

Scale-Invariant Framework: P-P Plots

The scale-invariant framework developed to address this pliability in PAC-based trends is based on the Probability-Probability (P-P) plot (Haertel, Thrash, & Wiley, 1978; Ho & Haertel, 2006a; Livingston, 2006; Spencer, 1983). A P-P plot is, "a comparative plot of sample cumulative probabilities" (Fisher, 1983, p. 31) "constructed solely from vertical slices across CDFs" (Ho, 2007, p. 13). Thus, P-P plots display the vertical gaps between CDFs of test scores administered at Time 1 and Time 2. But rather than focusing on a single vertical slice, such as the gap in proficiency, P-P plots display vertical gaps at all percentiles. As Ho (2007) notes, "a monotone transformation of scale may contort the CDFs horizontally, but will not change the vertical relationships between the cumulative proportions" (p. 13). Thus, P-P plots, and any statistics derived from them, are invariant to transformations of the score scale.

P-P curves, which increase monotonically from the origin to the point (1,1), display the percentiles of one distribution versus the percentiles of another distribution (Holmgren, 1995). When the distributions represent scores from the same test administered twice, the P-P curve shows the proportion of students scoring at or below a given cut-score at each time. In other words, for a given percent *p*,

$$F_2^{-1}(p_2) = \text{the } p^{\text{th}} \text{ percentile from Time 2} \quad (3)$$

(which represents the test score at which p% of students scored below at Time 2), the P-P plot displays

$$p_1 = F_1\left[F_2^{-1}(p_2)\right], \quad (4)$$

the percentage of students scoring at Time 1 scoring below given percentiles of Time 2. Figure 3.6 displays the P-P plot for the simulated data set from Figures 3.3 and 3.4.

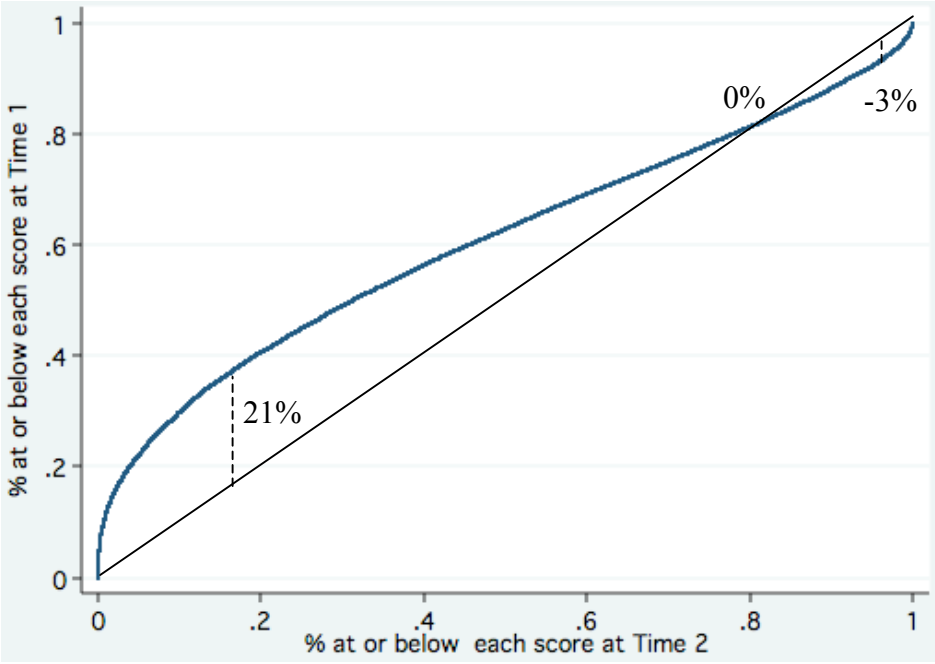


Figure 3.6 P-P plot from the simulated data displayed in Figures 3.3 and 3.4

The diagonal line in Figure 3.6 is shown for reference. A P-P function that lies on the diagonal would represent identical score distributions at Time 1 and Time 2, whereas a P-P function that lies mainly above the diagonal would indicate a positive score trend. The vertical lines drawn in Figure 3.3 show this. The point (.16, .37) on the P-P curve shows that only 16% of students at Time 2 scored below the 37th percentile from the Time 1 distribution (the same 21% "gain" displayed in Figures 3.4 and 3.5).

Summary Statistics: *V* Coefficients

The P-P plot is a scale-independent representation of the difference between two distributions, as horizontal scale distortions of the CDFs do not change the P-P plot. Statistics generated from P-P plots are likewise scale-invariant. Since vertical deviations from the P-P plots to the diagonal represent score trends, one useful and interpretable statistic of interest would be the area under the P-P curve.

The area under the P-P curve (AUC):

$$\text{AUC} = \int_0^1 F_1\left(F_2^{-1}\left(p_2\right)\right)dp_2 = P(X_2 > X_1), \quad (4)$$

represents the probability that a randomly chosen test score from the Time 2 distribution is greater than a randomly chosen test score from the Time 1 distribution (Ho, 2007, p. 14). For identical score distributions at Time 1 and Time 2, the area under the P-P curve (which would fall on the diagonal) would be 0.50 (representing the chance probability). When scores improve from Time 1 to Time 2, the P-P curve would fall above the diagonal and the area would be greater than 0.50. As Ho notes, "the usefulness of this statistic is that it is invariant to discretionary choices such as cut-scores, percentile, and score scale" (p. 8). Thus, the area under the P-P curve addresses the problem of pliability of PAC-based trend comparisons under choice of cut-scores.

For the P-P plot in Figure 3.6, the area under the curve is approximately 0.611. This positive value represents the positive trend in scores from Time 1 to Time 2. It also indicates that a randomly chosen test score from Time 2 has a 61% probability of being greater than a randomly chosen test score from Time 1.

Wilk and Gnanadesikan (1968) note that the nonparametric Kolmogorov-Smirnov statistic is represented as the maximum deviation from the 45-degree diagonal to a point on the P-P plot (p. 11). Likewise, Ho (2007) notes that P-P plots have conceptual ties to Receiver Operator Characteristic (ROC) curves and Lorenz Curves; and that the nonparametric Mann-Whitney U statistic is a simple linear transformation of the

$P(X_2 > X_1)$ statistic. Another useful transformation of $P(X_2 > X_1)$ is found by assuming that the distribution from Time 1 has a standard normal distribution and the distribution from Time 2 has a normal distribution with unit variance. Under these assumptions, the area under the P-P curve defines the mean for the distribution from Time 2 that can be interpreted in terms of standard deviation units. Thus, these assumptions lead to a transformed summary statistic:

$$V = \sqrt{2}\Phi^{-1}\left(P(X_2 > X_1)\right) = \sqrt{2}\Phi^{-1}\left(\int_0^1 F_1\left(F_2^{-1}\left(p_2\right)\right)dp_2\right), \quad (4)$$

where Φ^{-1} represents an inverse normal transformation. Ho and Haertel (2006a) describe V as a scale-free effect size of the trends in scores from Time 1 to 2. Unlike traditional effect sizes, the V statistic cannot be distorted by scale transformations, yet it may still be loosely interpreted as a distance in terms of standard deviation units.

For the P-P plot in Figure 3.6, $V \approx \sqrt{2}\Phi^{-1}(.611) \approx .40$. This indicates that the Time 2 scores increased by 0.40 standard deviation units over the Time 1 scores. This is supported by the fact that the distributions were simulated to have an effect size of approximately 0.40.

Calculating V From Reported Performance Level Data

P-P plots and V statistics are calculated from test score distributions. In this study, score distributions are not known. As Table 3.4 shows, the collected data simply show the percentage of students scoring at or below specific cut-scores for each state test and NAEP. If a state administers the same test twice and cut-scores do not change, then the corresponding percentages of students scoring below each cut-score at each time define points on a P-P plot.

As an example, consider the 2003-2005 8th grade math (highlighted) data reported in Table 3.4. The data show that from 2003-2005, the percentage of students scoring below the first cut-score on the state test decreased from 33.2% to 25.3%. From this

information, the point (.253, .332) was placed on the P-P plot. Likewise, the points (.703, .793) and (.941, .977) were plotted on the P-P curve for the state test. For the NAEP data, the points (.330, .306), (.753, .758), and (.981, .983) were plotted.

Table 3.4 Example of collected test data to interpolate P-P curves

South Carolina	State Test			NAEP		
	2003	2005	2007	2003	2005	2007
4th Grade Reading	18.9	21.4	21.9	20.8	18.5	20.3
	65.6	59.5	58.6	68.2	64.1	64.2
	86.0	85.8	80.3	96.1	95.3	95.3
4th Grade Math	23.6	20.4	17.3	40.6	42.6	41.1
	67.1	63.6	57.8	74.3	74.4	74.2
	97.7	97.1	95.9	94.7	94.2	94.6
8th Grade Reading	32.9	33.7	32.1	32.2	28.6	29.1
	80.2	76.8	80.2	73.7	70.1	68.1
	93.7	92.0	93.2	95.2	93.3	92.6
8th Grade Math	*33.2*	*25.3*	28.7	*30.6*	*33.0*	31.3
	79.3	*70.3*	75.4	*75.8*	*75.3*	75.4
	97.7	*94.1*	96.6	*98.3*	*98.1*	98.3

Notes: Numbers represent percentages of students scoring below 3 cut-scores.

Thus, states reporting data from at least 3 cut-scores provided 3 points for the P-P plot. The theoretical points (0, 0) and (1, 1) were then added to the P-P plot to yield five data points. Figure 3.7 displays these points plotted for the example data in Table 3.4.

Using these five points, the smoothed curve algorithm implemented in Microsoft Excel was then used to plot a curve from a cubic Bezier-based interpolation function with control points (Ho, 2007). Figure 3.7 displays the smoothed (interpolated) P-P curves for the example data in Table 3.4. According to Ho and Haertel (2006), "Simulation studies suggest that three P-P points is the minimum number of points necessary for the interpolation function to obtain a reasonable approximation of the P-P curve" (p. 34). Thus, data from states reporting fewer than three cut-scores were eliminated from this study

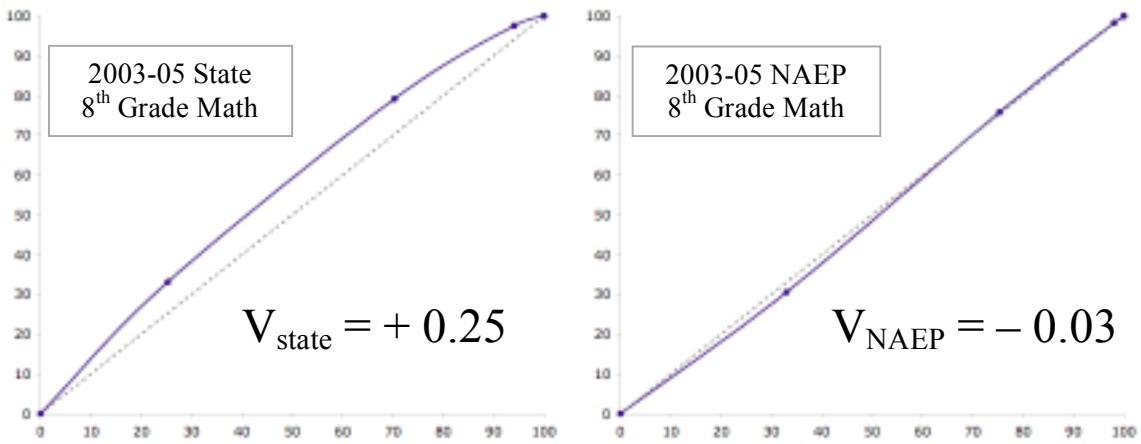


Figure 3.7 Smoothed (interpolated) P-P plots for the example data in Table 3.4.

The cubic spline macro for Microsoft Excel (SRS1 Software, 2007) was then used to obtain interpolated points from this smoothed P-P curve. With these points, numerical integration procedures were used to estimate the area under the P-P curve. From these estimated areas, values of the V statistic were calculated. In this study, Simpson's Rule was used to estimate the area under the smoothed P-P curves using 10,000 interpolated subdivisions. These area estimates were identical (to at least 6 decimal places) to the area estimates using 50,000 interpolated subdivisions.

Applying these procedures to the example data in Figure 3.7, V statistic values were estimated to be 0.25 for the state test and -0.03 for NAEP. Thus, for 8th grade math in South Carolina from 2003-2005, state test trends indicate an increase of 0.25 standard deviation units, while NAEP trends indicate a decline of 0.03 standard deviation units. Thus, the discrepancy in score trends was estimated to be $V = 0.28$.

Since all test scores contain measurement error, the effect of this error on the scale invariant effect sizes is of concern. As Ho (2007) explains:

> Effect size-based trend statistics are generally attenuated by measurement error, but NAEP reports statistics that are corrected

> for this effect. In contrast, [state test] effect sizes are biased towards zero due to measurement error. If [state test] V statistics are treated like traditional effect sizes, they can be corrected by disattenuating them in inverse proportion to the square root of the reliability of the test (Hedges & Olkin, 1985). As reliabilities for State assessments are not always reported, the uncorrected [state test] statistics are used. As the results will show, if the reliabilities of state tests are taken into account, disattenuation will increase the degree of average State-NAEP trend discrepancy (p. 16).

Final Sample of Test Score Data

Tables 3.5 and 3.6 display the final sample of state test data that was included in, or excluded from, analysis. For each state, the scale-invariant effect size V for trends in state test and NAEP results were computed. Recall that, ideally, the data set would include results from each state in 2003, 2005, and 2007 in reading and mathematics in grades 4 and 8 for a total of 600 state test trend effect sizes and 600 NAEP trend effect sizes. Due to changes in tests or cut-scores, a lack of available data, or the use of fewer than three cut-scores, the final data set included 215 state test score distributions.

Once the data had been entered, the P-P plots generated through the interpolation function were visually inspected to check the accuracy of data entry. P-P plots with extreme deviations from the diagonal or that cross the diagonal were flagged for further investigation. Estimated values of AUC and V were also inspected and outliers were checked again for accuracy.

Visual displays of the V estimates were also examined to determine the accuracy of data entry and analysis. Since the V estimates should tend to agree with the proficiency trends reported by states, a scatterplot was inspected. Any observations in which the V estimates and proficiency trends differ were flagged for further investigation. A scatterplot of the V estimates and traditional effect sizes (calculated from the states that report means and standard deviations) were also inspected and outliers were flagged. A similar scatterplot using NAEP V estimates and traditional effect sizes was likewise inspected to determine the accuracy of data entry and analysis.

Table 3.5 Data included in the analysis

	Number of cut-scores reported on state tests											
	Grade 4						Grade 8					
	Reading			Mathematics			Reading			Mathematics		
	03-05	05-07	03-07	03-05	05-07	03-07	03-05	05-07	03-07	03-05	05-07	03-07
Alabama		3	3		3	3		3	3		3	3
Alaska												
Arizona	3	3	3	3	3	3	3	3	3	3	3	3
Arkansas		3			3			3			3	
California	4	4	4	4	4	4	4	4	4	4		
Colorado	3	3	3		3		3	3	3	3	3	3
Connecticut	4			4			4			4		
Delaware							4	4	4	4	4	4
Florida	4	4	4	4	4	4	4	4	4	4	4	4
Georgia												
Hawaii		3			3		3	3	3	3	3	3
Idaho	3	3	3	3	3	3	3	3	3	3	3	3
Illinois							3			3		
Indiana												
Iowa												
Kansas				4	4	4	4	4	4			
Kentucky		3									3	
Louisiana	4	4	4	4	4	4	4	4	4	4	4	4
Maine	3	3	3	3	3	3	3	3	3	3	3	3
Maryland												
Massachusetts		3			3						3	
Michigan	3	3	3	3	3	3				3	3	3
Minnesota												
Mississippi	3	3	3	3	3	3	3	3	3	3	3	3
Missouri				3						3		
Montana	(3)	(3), 3	3	(3)	(3), 3	3	(3)	(3), 3	3	(3)	(3), 3	3
Nebraska												
Nevada												
New Hampshire												
New Jersey												
New Mexico												
New York	3			3			3			3		
North Carolina	3			3			3			3		
North Dakota		3			3			3			3	
Ohio		4						4			4	
Oklahoma		3			3		3	3	3	3	3	3
Oregon												
Pennsylvania							3	3	3	3	3	3
Rhode Island												
South Carolina	3	3	3	3	3	3	3	3	3	3	3	3
South Dakota												
Tennessee												
Texas	4			4			4			4		
Utah												
Vermont												
Virginia												
Washington	3	3	3	3	3	3						
West Virginia		4			4			4			4	
Wisconsin	3	3	3	3	3	3	3	3	3	3	3	3
Wyoming	3			3			3			3		

Notes: Blank cells represent data excluded from the analysis

Montana administered 2 tests in 2003 & 2005. The (ITBS) changed from high- to low-stakes in 2005.

Table 3.6 Data excluded from the analysis

	Number of cut-scores reported on state tests											
	Grade 4						Grade 8					
	Reading			Mathematics			Reading			Mathematics		
	03-05	05-07	03-07	03-05	05-07	03-07	03-05	05-07	03-07	03-05	05-07	03-07
Alabama	T			T			T			T		
Alaska	TS	C	TCS	TS	C	TCS	TS	C	TCS	TS	C	TCS
Arizona												
Arkansas	T		T	T		T	T		T	T		T
California												
Colorado				D		D						
Connecticut		S	S		S	S		S	S		S	S
Delaware	D	D	D	D	D	D						
Florida												
Georgia	C	C	C	C	C	C	C	C	C	C	C	C
Hawaii	D		D	D		D						
Idaho												
Illinois	D	D	D	D	D	D		D	D		D	D
Indiana	CD	C	CD	CD	C	CD	C	C	C	C	C	C
Iowa	C	C	C	C	C	C	C	C	C	C	C	C
Kansas	D	D	D							D	D	D
Kentucky	S		S	D	D	D	D	D	D	S		S
Louisiana												
Maine												
Maryland	D	C	CD	D	C	CD	C	C	C	C	C	C
Massachusetts	D		D	D		D	D	D	D	D		D
Michigan							D	D	D			
Minnesota	D	D	D	D	D	D	D	D	D	D	D	D
Mississippi												
Missouri	D	TS	TSD		TS	TS	D	TSD	TSD		TS	TS
Montana												
Nebraska	D	D	D	D	D	D	D	D	D	D	D	D
Nevada	D	D	D	D	D	D	D	D	D	D	D	D
New Hampshire	D	D	D	D	D	D	D	D	D	D	D	D
New Jersey	C	CD	CD	C	CD	CD	C	CD	CD	C	CD	CD
New Mexico	T	D	T	T	D	T	T	D	T	T	D	T
New York		D	D		D	D		D	D		D	D
North Carolina		D	D		D	D		D	D		D	D
North Dakota	S		S	S		S	S		S	S		S
Ohio	T		T	T	D	T	D		D	D		D
Oklahoma	T		T	T		T						
Oregon	C	CS	CS	C	CS	CS	C	CS	CS	C	CS	CS
Pennsylvania	D	D	D	D	D	D						
Rhode Island	T	T	T	T	T	T	T	T	T	T	T	T
South Carolina												
South Dakota	C	C	C	C	C	C	C	C	C	C	C	C
Tennessee	T	C	TC	T	C	TC	T	C	TC	T	C	TC
Texas		C	C		C	C		C	C		C	C
Utah	C	C	C	C	C	C	C	C	C	C	C	C
Vermont	D	D	D	D	D	D	D	D	D	D	D	D
Virginia	D	D	D	D	D	D	C	C	C	C	C	C
Washington							D	D	D	D	D	D
West Virginia	T		T	T		T	T		T	T		T
Wisconsin												
Wyoming		T	T		T	T		T	T		T	T

Notes: Blank cells represent data that were included in the analysis.

T = data excluded due to change in state tests; C = data excluded due to state reporting fewer than 3 cut-scores

S = data excluded due to change in scoring standards (or number of cut-scores reported); D = data not reported

Reporting

Once values of *V* are estimated for state test and NAEP trends, the values for each state are displayed on a scatterplot to show the discrepancies. The centroid of the scatterplot represents the average discrepancy between state test and NAEP trends. The axes for the scatterplot are displayed in Figure 3.8. Because the values of *V* arrive from a normality assumption, a matched-pairs t-test is used to test the null hypothesis of equal trends in both state and NAEP tests.

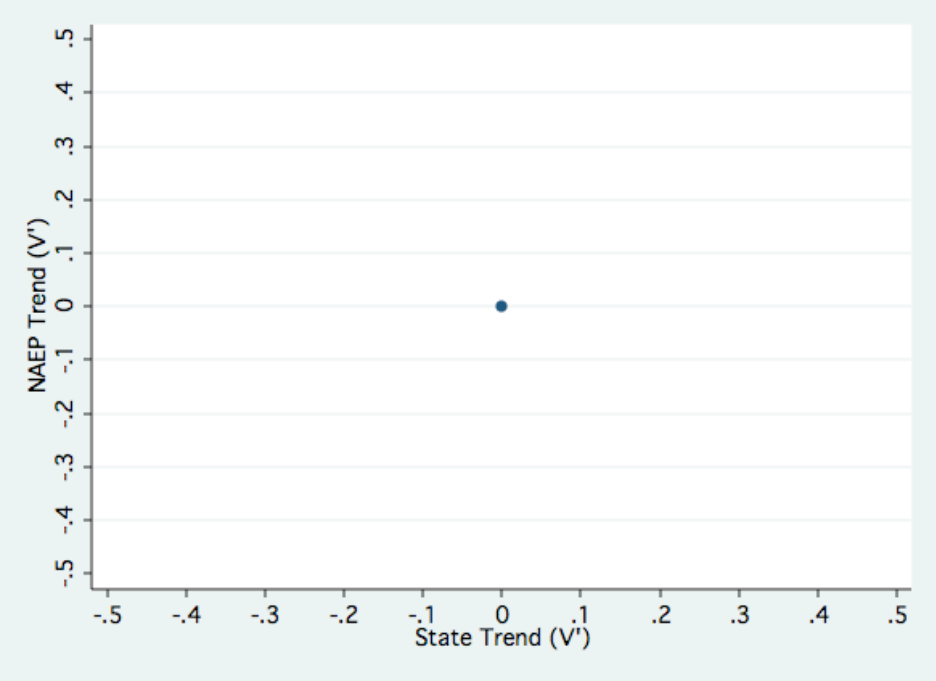


Figure 3.8 Scatterplot to display scale-invariant trend effect sizes

With a sample size of 215, the power of this t-test can be estimated via G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) by making some assumptions. Based on the 2003-05 discrepancy results reported by Ho (2007), assume the true discrepancy between state and NAEP trend effect sizes is 0.80 with a standard deviation of 0.15 for state

trends, a standard deviation of 0.10 for NAEP trends, and a correlation of 0.60 between state and NAEP trends. From these assumptions, a one-tailed t-test with an alpha of 0.01 would have estimated power above 0.99. These assumptions lead to an assumed effect size of 0.70 in the difference between state test and NAEP trends. With a sample size of 215, an effect size of at least 0.25 will be required for power to be 0.90. If alpha is set at 0.05, then an effect size of at least 0.20 will lead to a power above 0.90.

To further summarize the discrepancies in score trends, the percentages of paired observations (dots on the scatterplot) falling above and below the 45-degree diagonal are reported. Dots below the diagonal represent state-subject-grade-year combinations in which state score trends were more positive than NAEP trends. The percentages of dots falling in each quadrant of the scatterplot are also reported to show how many state-subject-grade-year trend combinations show a difference in sign (positive state trends with negative NAEP trends or vice versa).

The above analyses are also computed for each subject, each grade level, and each pair of years (2003-05, 2005-07, and 2003-07) to see if any different conclusions are reached. Finally, for each state-subject-grade-year combination, simple differences in V values are computed. These differences in V values will be used as a single estimate of the discrepancy between state test and NAEP results.

Assumptions and Limitations

One methodological limitation of this study involves the interpolation function used to generate the P-P plots from reported state test data. P-P plots and values of V are estimated from the limited number of cut-scores reported by each state. The estimations for states with fewer cut-scores probably contain more error than states that report larger numbers of cut-scores. Ho and Haertel (2006) note that, "the cubic Bezier interpolation method provides an estimate whose error has not been assessed" (p. 39). The interpolation procedure also requires the elimination of data from any state test with

fewer than 3 cut-scores. The elimination of this data may bias the results in some unknown way.

An analysis was conducted to investigate the impact of the choice of 3 cut-scores as a limitation of this method. For this analysis, the V estimates for states reporting 4 cut-scores were set aside. Then, for each of these states, the first cut-score was eliminated and V was re-estimated using the remaining 3 cut-scores. Likewise, V was estimated from 3 cut-scores after eliminating the second, third, and fourth reported cut-score. Comparisons were then made among the V estimates from 4 and 3 cut-scores. A significant discrepancy in the estimates would make the choice of 3 cut-scores suspect. Even if no significant discrepancy is found, the results of this study depend, in small part, to the interpolation method chosen to construct the smoothed P-P curves.

Another possible limitation is that V estimates are calculated by applying a normal transformation to the area under the P-P curve estimates. The value of these V estimates would change under different distributional transformations. The results in this study, therefore, are dependent upon this choice of transformation to normal distributions.

The usefulness of the V estimates was also analyzed through comparisons with traditional effect sizes (the ratio of the mean difference to the pooled standard deviation) from states reporting score means and standard deviations. Although this resulted in a small sample of data, it provides evidence as to the usefulness of the V estimates. To do this, a scatterplot (and the standard error of estimate) of the relationship between V and traditional effect size estimates were examined. Outliers were investigated in an attempt to provide an explanation as to why the V estimate would differ from the traditional effect size estimate.

The quality of the data also could have negatively impacted this study. The estimated NAEP percentiles and published state test results (which are oftentimes subject to rounding) contain error that impacted the V estimates. Also, some states only reported English or Language Arts test scores instead of reading test scores (as is reported

by NAEP). Trend discrepancies in these cases could simply be due to significant content differences in the tests. Also, some states administered their tests in the fall, whereas the NAEP is administered in the spring. This could have impacted results, since fall-to-fall score trends would only be similar to spring-to-spring score trends if student achievement grows at a nearly linear rate.

Analysis

As explained earlier, the data collected from this analysis consists of estimates of the discrepancy in trends between state tests and the NAEP ($V_{state} - V_{NAEP}$) and various measures of the quality of state test security policies. From this data, the following questions can be addressed:

- What is the relationship between the quality of a state's test security policy and any discrepancies between score trends on state and NAEP tests? This question is addressed by regressing ($V_{state} - V_{NAEP}$) on the overall security policy composite score. This analysis is conducted separately for each grade level and subject.
- Which aspects of a state's test security policy seem to have the strongest relationship with score trend discrepancies? This question is addressed by regressing ($V_{state} - V_{NAEP}$) on composite scores for each of the four components (F-O-I-L) of the policy evaluation form. A comparison of the standardized beta weights would indicate which aspect had the greatest impact on trend discrepancies.
- Do states with higher numbers of reported test score manipulations experience greater trend discrepancies? To address this question, the correlation between the numbers of published news reports and the trend discrepancies for each state is calculated. An analysis is conducted to determine if states with higher numbers of published reports prior to 2003 experienced smaller discrepancies (evidence that the published reports caused states to focus more on test security).

The seven dichotomizations of state policies are used to conduct mean comparison tests in order to address the following questions:

- Do states with clear and accessible policies experience smaller trend discrepancies than states with ambiguous or difficult-to-find policy information? If discrepancies did reflect manipulations, then states with clear policies would be expected to have smaller discrepancies than states with ambiguous policies.
- Do states with state-level mandates experience smaller trend discrepancies than states with district-level policies? State policies might carry more weight with regards to sanctions, but district policies might be better implemented.
- Do states with punitive policies experience smaller trend discrepancies than states with instructive policies? This comparison could provide some information to determine if sanctions were more effective than honor codes in deterring manipulations.
- Do states requiring independent test administration monitoring experience smaller trend discrepancies than states that allow teachers to administer tests to their own students? Independent monitoring would make it more difficult to manipulate test scores.
- Do states with investigative policies experience smaller trend discrepancies than states with preventative policies? Investigative policies focus on mandatory reporting of testing irregularities and the process of investigating those reports. Preventative policies focus more on informing educators about which activities and behaviors are appropriate and inappropriate.
- Do states with example-based policies experience smaller trend discrepancies than states with more general policies? It is hypothesized that policies with more examples would be more effective than policies with fewer examples.
- Do states with positive-message policies experience smaller trend discrepancies than states with negative-message policies? This could provide information to

determine if policies designed to "scare" educators from manipulating scores are more effective than policies designed as honor codes.

After these analyses, all possible longitudinal analyses for states that adopted or modified test security policies between 2005-2007 were investigated. Unfortunately, while many states did update their policies during this period, the magnitude of the changes could not always be determined. Thus, conclusions from the longitudinal analyses must be interpreted carefully.

Scope

This study analyzes data from 2003, 2005, and 2007 for grades 4 and 8 in reading and mathematics in states with tests having at least 3 cut-scores. It only analyzes results from state tests and the NAEP in those years, subjects, and grades. Conclusions are cautiously made due to the presence of confounding variables.

Assumptions, Limitations, and Confounding Variables

The biggest limitation of this study is that it cannot attempt to find a causal relationship between test security policy quality and score trend discrepancies. A causal relationship cannot be inferred due to the influence of confounding variables (plausible rival hypotheses, discussed in the previous chapter). Differences in test content, test administration, examinee motivation, examinee populations, and other factors could influence the relationship between the independent and dependent variables in this study. Likewise, this study can only evaluate the quality of test security policy materials. No causality can be concluded because the quality of a state's policy materials does not necessarily represent the quality of the state's policy implementation.

For example, one alternate hypothesis for any significant relationship between test security policy quality and trend discrepancies is that states with higher-quality policies simply have higher-quality testing programs. Higher-quality testing programs could, in turn, simply represent states with higher-quality educational systems (curriculum

development, teacher training, etc.). States with higher-quality educational systems might be expected to score higher on state tests (trained teachers focusing on state-developed curriculum) than on NAEP. This is just one of many plausible hypotheses. This study only attempts to discover if a relationship exists between the quality of state test security policies and score trend discrepancies.

A second limitation is that some states were excluded from analysis. Data from states experiencing changes in tests or cut-scores, or from states with fewer than three cut-scores, were excluded. The exclusion of this data may bias the results in some unknown way and limit the generalization of the results.

A third limitation is that the trend discrepancy statistics are calculated at the state-level. The decision to manipulate test scores may be made at the level of individual teachers, schools, or school districts. If, in fact, manipulations have a relationship with score trend discrepancies, a state-level trend estimate may not be able to detect this relationship.

Summary

This study has been designed to determine if a relationship exists between the existence and quality of state test security policies and discrepancies in state-NAEP score trends. The quality of state test security policies is evaluated based on a framework derived from analyses of newspaper reports, educator surveys, state surveys, direct observation studies, statistical analyses, and targeted research into test score manipulations. State-NAEP trend effect size discrepancies are estimated via a scale-invariant framework that is not impacted by choice of cut-scores.

The study is designed in such a way as to determine which aspects of test security policies might be more strongly related to score trend discrepancies. The study also provides information regarding the content and quality of test security policies adopted by states to deter educators from manipulating scores.

The study extends the work of researchers who have focused on inappropriate testing practices, test score pollution, and detecting cheating on achievement tests. It also extends the work of researchers who have focused on comparing state test results to NAEP results. It extends the work of Ho (2005, 2007) and Haertel (2006a) in their development of scale-invariant measures of trend discrepancies and furthers the research of Cizek (1999) and McCabe and Trevino (1993, 2002) in evaluating the impact of test security practices and honor codes in educational organizations.

Results of this research could be used by states to develop, improve, or audit their current test security policies and practices. Results could also be used in professional development to train teachers in appropriate test preparation and administration activities. Finally, this study will contribute to the debate over the effectiveness of accountability systems and sanctions in improving student achievement.

CHAPTER 4: RESULTS AND ANALYSIS

To fulfill the purpose of this study, the quality of state test security policies is first evaluated. Then, state-NAEP trend discrepancies are estimated using the scale-invariant framework discussed previously. Finally, this information is used to determine if a relationship exists between the quality of state test security policies and discrepancies between state test and NAEP score trends.

Quality of State Test Security Policies

The FOIL framework discussed previously was used to evaluate the quality of state test security policy materials in an attempt to answer the following research questions:

- What test security policies and practices do states implement in an attempt to deter educators from manipulating test scores?
- What is the quality of each state's test security policy?

Evaluation forms were completed for each state. To summarize the information, state composite scores are first reported. Then, scores for each component in the framework are discussed. These components are then used to classify states in an attempt to discuss state policies in greater detail.

Table 4.1 displays the composite scores for the F(ormalize), O(versee), I(nform), and L(imit) components for each state. The composite scores represent simple sums of the individual ratings under each component. The table shows that, overall, Michigan had the highest-rated policy, earning 103 out of 122 (84%) possible points on the evaluation scale. Texas, Illinois, Kentucky, and Wisconsin had the next-highest scores. On the other end of the spectrum, policies from Iowa, Nebraska, Maine, Missouri, and New Jersey earned the lowest composite scores, with Iowa's policy earning a score of 4 out of 122 (3%).

Table 4.1 Policy evaluation scores for each state

	Formalize (48 max)	Oversee (18 max)	Inform (44 max)	Limit (12 max)	Composite (122 max)
Michigan*	**42 (88)**	**12 (67)**	**41 (93)**	8 (67)	**103 (84)**
Texas*	**43 (90)**	**14 (78)**	29 (66)	**10 (83)**	**96 (79)**
Illinois*	35 (73)	**10 (56)**	32 (73)	**12 (100)**	**89 (73)**
Kentucky*	**41 (85)**	7 (39)	29 (66)	8 (67)	**85 (70)**
Wisconsin*	32 (67)	8 (44)	**36 (82)**	8 (67)	**84 (69)**
North Carolina*	**37 (77)**	3 (17)	**33 (75)**	**10 (83)**	83 (68)
South Carolina*	36 (75)	8 (44)	30 (68)	6 (50)	80 (66)
Delaware*	**37 (77)**	7 (39)	27 (61)	8 (67)	79 (65)
Arizona*	35 (73)	3 (17)	30 (68)	7 (58)	75 (61)
Minnesota	**38 (79)**	2 (11)	30 (68)	4 (33)	74 (61)
Utah	32 (67)	2 (11)	**33 (75)**	7 (58)	74 (61)
Louisiana*	34 (71)	**13 (72)**	20 (45)	6 (50)	73 (60)
Mississippi*	28 (58)	**14 (78)**	20 (45)	6 (50)	68 (56)
Oklahoma*	27 (56)	2 (11)	32 (73)	6 (50)	67 (55)
Florida*	31 (65)	2 (11)	27 (61)	6 (50)	66 (54)
Washington*	27 (56)	1 (6)	**33 (75)**	5 (42)	66 (54)
Nevada	36 (75)	3 (17)	20 (45)	5 (42)	64 (52)
New Mexico	26 (54)	2 (11)	29 (66)	7 (58)	64 (52)
Pennsylvania*	30 (63)	4 (22)	23 (52)	6 (50)	63 (52)
West Virginia*	29 (60)	4 (22)	22 (50)	7 (58)	62 (51)
Virginia	29 (60)	0 (0)	23 (52)	**9 (75)**	61 (50)
Ohio*	30 (63)	1 (6)	24 (55)	5 (42)	60 (49)
Georgia	29 (60)	0 (0)	21 (48)	6 (50)	56 (46)
South Dakota	21 (44)	2 (11)	25 (57)	7 (58)	55 (45)
California*	23 (48)	4 (22)	22 (50)	5 (42)	54 (44)
Montana*	29 (60)	1 (6)	21 (48)	3 (25)	54 (44)
Tennessee	21 (44)	4 (22)	22 (50)	6 (50)	53 (43)
Alaska	21 (44)	1 (6)	23 (52)	6 (50)	51 (42)
North Dakota*	19 (40)	0 (0)	26 (59)	6 (50)	51 (42)
Colorado*	15 (31)	0 (0)	27 (61)	8 (67)	50 (41)
Oregon	22 (46)	0 (0)	22 (50)	5 (42)	49 (40)
Hawaii*	22 (46)	0 (0)	18 (41)	7 (58)	47 (39)
Vermont	18 (38)	0 (0)	23 (52)	5 (42)	46 (38)
Indiana	18 (38)	0 (0)	22 (50)	4 (33)	44 (36)
Arkansas*	17 (35)	0 (0)	17 (39)	8 (67)	42 (34)
Idaho*	14 (29)	0 (0)	20 (45)	8 (67)	42 (34)
Massachusetts*	18 (38)	1 (6)	17 (39)	4 (33)	40 (33)
Maryland	19 (40)	4 (22)	14 (32)	2 (17)	39 (32)
New York*	18 (38)	0 (0)	15 (34)	6 (50)	39 (32)
Wyoming*	11 (23)	0 (0)	23 (52)	4 (33)	38 (31)
Connecticut*	19 (40)	0 (0)	13 (30)	5 (42)	37 (30)
Alabama*	14 (29)	1 (6)	17 (39)	3 (25)	35 (29)
Kansas*	10 (21)	0 (0)	24 (55)	0 (0)	34 (28)
New Hampshire	13 (27)	0 (0)	6 (14)	5 (42)	24 (20)
Rhode Island	13 (27)	0 (0)	6 (14)	5 (42)	24 (20)
New Jersey	6 (13)	0 (0)	7 (16)	6 (50)	19 (16)
Missouri	7 (15)	1 (6)	8 (18)	2 (17)	18 (15)
Maine*	9 (19)	0 (0)	8 (18)	0 (0)	17 (14)
Nebraska	3 (6)	2 (11)	3 (7)	0 (0)	8 (7)
Iowa	0 (0)	0 (0)	4 (9)	0 (0)	4 (3)

Notes: Bold values represent the five highest scores in each column

(Values in parentheses represent percentage of total points in each component)

* represents data included in the complete study (discrepancies analysis)

The state policies with the highest composite scores also tended to have the highest scores in each component. The only exceptions were Minnesota with a high score in the Formalize component, Louisiana and Mississippi with high Oversee component scores; Utah and Washington with high scores in the Inform component; and Virginia having a high score in the Limit component. To investigate the relationships among the components, Table 4.2 displays the Spearman rank-order correlations among the components and the composite scores.

Table 4.2 Spearman rank-order correlations between policy component scores

	Formalize	Oversee	Inform	Limit	Composite
Formalize	1.00				
Oversee	0.72	1.00			
Inform	0.72	0.47	1.00		
Limit	0.56	0.35	0.56	1.00	
Composite	0.95	0.74	0.84	0.67	1.00

The table shows that while all four components and the composite score have moderate positive correlations, they do seem to measure something unique in the policies. The strongest inter-component correlations were found between the Formalize and Oversee and Formalize and Inform components. The smallest correlation was found between the Oversee and Limit components. These relationships are also displayed in the scatterplot matrix of Figure 4.1.

Before taking a more in-depth look at the scores for each component and state, one important note from Table 4.1 must be mentioned. As discussed earlier, the policy evaluation ratings will be compared with estimates of the discrepancies between state test and NAEP results to determine if a relationship exists. Some states will be excluded from this analysis because of a lack of data to estimate discrepancies. The states that will be included in this analysis are highlighted (*) in Table 4.1. Looking at which states are

highlighted, one can see that 9 of the 10 states with the highest evaluation scores will be included in the study, but only 4 out of the 10 lowest scoring states will be included. In fact, only 13 of the bottom 25 states will be included in the analysis. The fact that the analysis will contain more states with higher policy scores than low scores may have an impact on the results.

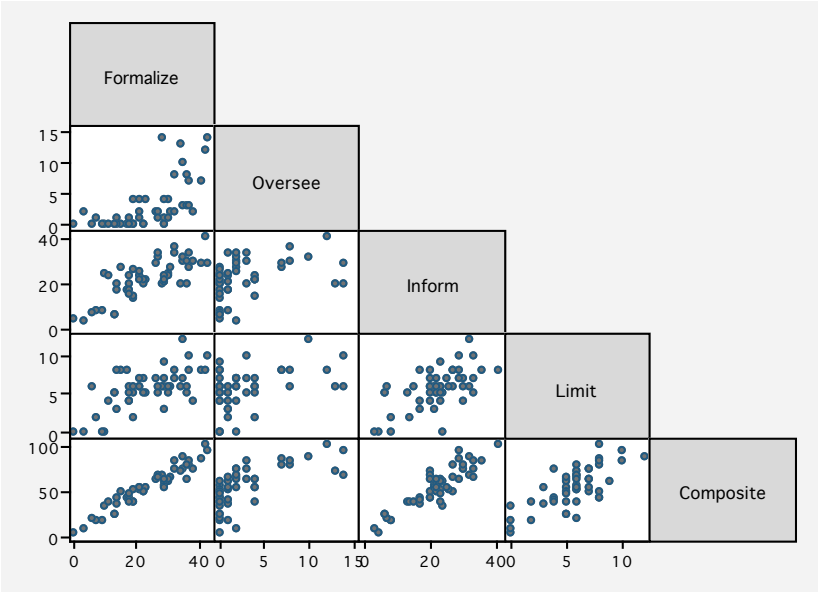


Figure 4.1 Scatterplot to display scale-invariant trend effect sizes

Policy Evaluation Component Scores

Table 4.3 displays the minimum, median, and maximum points earned by states in each component along with a summary of other information collected from state policies. For comparison, the maximum possible points in each component are also displayed. Distributions of the component scores are displayed in Figure 4.2. For example, Figure 4.2a shows that the distribution of composite scores for state policies ranged from 4 to 103, with a median of 54 (44% of the maximum possible points). These distributions provide information for a more detailed discussion of each policy component score.

Table 4.3 Summary of policy evaluation ratings

	Minimum	Median	Maximum	Max Possible
Formalize beliefs about the role of testing & testing practices	0	22.5	43	48
Prominence / Availability of information	0	4	10	10
Content	0	5	10	10
Implementation	0	7.5	12	12
Requirements and sanctions	0	5	12	12
Other	0	2	4	4
Oversee test preparation, administration, and scoring activities	0	1	14	18
Test Security Audits	0	0	2	2
Test administration oversight	0	0	3	4
Statistical Analyses	0	0	10	10
Score Reports	0	0	2	2
Inform educators about why some activities are unacceptable	3	22.5	41	44
Principles & Rules	0	5	12	12
Examples of appropriate and inappropriate behaviors	0	4	9	10
General Guidance	0	7	12	12
Training	0	6	10	10
Limit opportunities for educators to manipulate test scores	0	6	12	12
Materials security	0	5.5	8	8
Test Forms	0	0	4	4
COMPOSITE SCORE	**4**	**54**	**103**	**122**

Number of policy documents available	1	10	30
Number of clicks to navigate to security page	2	4	6
% of test administration manual pages dedicated to security	0%	4%	23%
Number of examples of appropriate and inappropriate activities	0	48	135
Number of appropriate examples (excluding accommodations)	0	6	44
Number of inappropriate examples (excluding accommodations)	0	8	49
Number of accommodations examples	0	25	100

Number of states with the following:		
# of states with positive explanation for security importance	17	34%
# of states with negative explanation for security importance	11	22%
Clear and accessible	30	60%
Ambiguous or difficult-to-find	19	38%
State-level mandate	22	44%
District- or school-level responsibility	18	36%
State and district responsibility (equally shared)	6	12%
Punitive or law-focused	15	30%
Instructive/informative	23	46%
Independent monitoring	11	22%
No independent monitoring	39	78%
Investigative	17	34%
Preventative	23	46%
Example-based	21	42%
Not many examples	28	56%
Positive message	20	40%
Negative message	17	34%

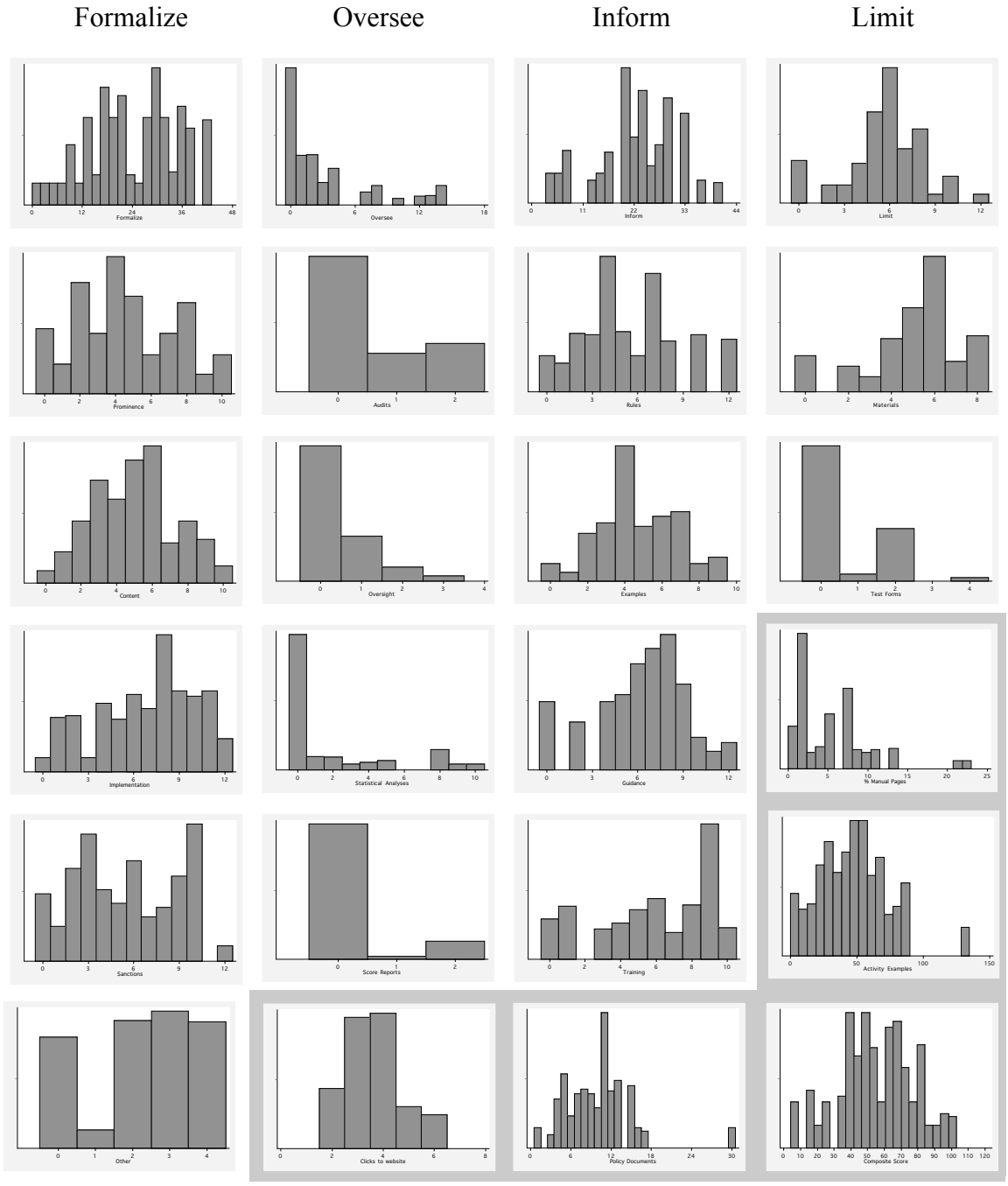


Figure 4.2 Distributions of security policy evaluation ratings

Component Score: Formalize

The first component score evaluated each state's policy with regard to how well it *formalizes* the beliefs of state educators regarding the role of testing and testing practices. As Table 4.3 and the first column of Figure 4.2 show, state policies varied quite a bit on this component. With 48 points possible, scores on this component ranged from 0 to 43 with a median of 22.5 (47% of the maximum possible). Table 4.4 displays more information about this component, including the number of states receiving scores of 0, 1, or 2 for each subcomponent. The subcomponents listed in Table 4.4 are ordered by the relative strength of the average state policy.

Table 4.4 shows that within the Formalize component, the highest average state policy score was earned in *implementation* (the average policy earned 57% of the total possible points in the implementation subcomponent). The *implementation* score was calculated from 6 items. States, on average, earned the highest score in identifying individuals, at both the state and district level, who are responsible for test security. For this item, 19 states identified individuals at both the state and district levels; 23 states identified individuals at either the state or district levels (but not both); and only 8 states failed to identify any individuals responsible for security. Most states also encouraged implementation by providing standard security forms and checklists for districts and schools to use, with 20 states providing at least one form, 20 states providing more than one form, and 10 states not providing any forms.

States did not fare quite as well in the evaluation of how regularly state policies are updated. While 21 states provided evidence that policies are updated at least every three years, 13 states had not updated their security policies in over a decade. Another opportunity for states to improve the implementation of their policies is in providing systems to automate the identification of students and their demographic information. While 21 states provide barcodes to automate the demographic coding on tests, 16 states

had optional barcodes and 13 states provided no automated system. Finally, states have an opportunity to improve security by collecting evidence that districts and schools are implementing the policies. 20 states have no system to judge if districts are implementing their test security policies.

Table 4.4 Formalize subcomponent distributions

Formalize beliefs about the role of testing & testing practices	Number of states earning 0	1	2	Avg.
• Implementation	Mean: 6.86 (57% of max)			
IDs individuals responsible for security at state and district levels	8	23	19	1.22
Availability of forms and checklists for districts/schools to use	10	20	20	1.20
Identifies an individual or group in charge of security	12	17	21	1.18
Policy content is updated regularly	13	16	21	1.16
Policy provides for a system to automate test score identification	13	16	21	1.16
Evidence of a dissemination plan being followed	20	13	17	0.94
• Other	Mean: 2.20 (55% of max)			
Laws, regulations, rules are mentioned	15	5	30	1.30
Explains the importance of test security	22	11	17	0.90
• Policy content	Mean: 5.00 (50% of max)			
Clarity of test security policy information	6	19	25	1.38
Rating of security policy documents	1	29	20	1.38
The policy requires educator signatures to indicate understanding	17	7	26	1.18
Availability of FAQs regarding test security	31	7	12	0.62
Teachers provided input into content	33	12	5	0.44
Number of policy documents available	--	--	--	9.86
• Requirements and sanctions	Mean: 5.30 (44% of max)			
Mandatory reporting requirements (for suspected incidents)	7	13	30	1.46
Sanctions for confirmed cases of manipulation are outlined	17	9	24	1.14
Sanctions include suspension/dismissal of confirmed manipulators	20	5	25	1.10
Due process is explained (procedures for investigating incidents)	26	8	16	0.80
Provides standard forms/online reporting for suspected incidents	32	4	14	0.64
Explains protections for individuals who report suspected incidents	45	2	3	0.16
• Prominence / availability of information	Mean: 4.32 (43% of max)			
The policy is mentioned in test administration manuals	8	14	28	1.40
Rating for policy pages in test administration manuals	5	27	15	1.21
Rating for website	18	23	9	0.82
The state has a web page for information about test security	30	10	10	0.60
The state has a separate test security office or budget	39	4	7	0.36
% of administration manual pages dedicated to security	--	--	--	5.6%
# of clicks from homepage to security information on website	--	--	--	3.69

The next relative strength for states in formalizing their policies was in the *other* subcomponent. It's noted that 30 states mention state laws, rules, or regulations in their test security policies. While 15 states do not mention specific laws, rules, or regulations, all 50 state policies refer to general "state laws" in describing test security. One area for improvement is in explaining the importance of test security. While 17 state policies have a nice explanation of why security is important, 22 state policies do not even mention the importance of security. Furthermore, 39% of the state policies that do explain the importance of security explain it in a negative sense (such as, security is important to avoid sanctions).

The average state policy earned 50% of the maximum possible points in the *policy content* subcomponent. On average, state policies were supported by almost 10 documents (policy summaries, training modules, coordinator manuals, etc.). The documents were, for the most part, clear and accessible, with 25 states earning the highest score for policy clarity. States can improve the content of their policies by allowing teachers to provide input (only 17 states provided evidence that educators had input into the policies), providing answers to frequently asked questions in test security (only 19 states provided this service), and requiring educators to sign the policy document to indicate understanding (17 states did not require any signatures).

States also have opportunities to improve test security by focusing on reporting, investigating, and sanctioning educators who manipulate test scores. The average state policy earned 44% of the total possible points in the *requirements and sanctions* subcomponent. While 30 states have mandatory reporting of testing irregularities (and 13 states encourage reporting), only 18 states provide a standardized form for reporting these irregularities. Also, while more than half of all state policies outline sanctions for manipulating test scores (including suspension or dismissal), less than half of the state policies describe the process used to investigate irregularities and only 5 states provide

even a basic level of protection (or anonymity) for individuals reporting potential manipulations.

The average state policy earned 43% of the total possible points in the final subcomponent of formalizing their policies, *prominence or availability of information*. Most states seemed to understand that publishing test security policy information in a variety of places might lead to more secure testing. 42 states at least mentioned test security policies in test administration manuals and 20 states published web pages mentioning test security (10 states had web pages dedicated to security). On average, close to 6% of the pages in test administration manuals were dedicated to test security. To improve security, states may want to provide a separate office or budget to test security. Currently, only 11 states identified a test security office or committee and only 7 states provided evidence of money budgeted specifically for test security purposes.

With a score of 43, the policy from Texas rated highest in the formalize component. This is due, in large part, to the overwhelming number of policy documents available online. With 30 documents (updated regularly) addressing various components of test security, Texas has clearly formalized the beliefs of educators with regards to the role of testing and test practices. The policy documents, published in print and online by the state's office of test security, require signatures from all personnel involved in testing to ensure understanding. In fact, at least seven documents must be signed before testing begins. The documents identify individuals responsible for test security at all levels, explains state test security laws, and outlines sanctions placed against those who violate the policy. The mandatory reporting requirements seem to be working, too, as more than 800 testing irregularities were reported to the Department of Education during the 2005-06 school year (Cizek, 2005).

The way in which the policy from Texas stood apart from policies from other states was in the extent to which the policy is evaluated and updated regularly. In 2005, the Texas Education Agency commissioned a review of its test security policy and

procedures after allegations of security breaches at several schools. The review outlined 11 recommendations to improve test security within the state (Cizek, 2005) and all 11 recommendations were implemented by 2007 (Texas Education Agency, 2007b). Policy content was again updated in 2006 after statistical analyses indicated possible security breaches. While these constant updates indicate that the policy is reactionary, they do demonstrate that the Texas test security policy is updated regularly to align with current beliefs about testing practices and security.

Component Score: Oversee

The second component evaluated for each state's policy determined how well the policies provided oversight for test preparation, administration, and scoring activities. Table 4.3 shows that of 18 points possible in this component, the median state policy only earned 1 point (5.6%). Table 4.5 displays the subcomponent distributions to identify ways in which states can improve in this area.

While state policies did not fare well in this component, the relative strength of state policies in overseeing testing practices was in auditing the implementation of security policies at the district-level. 20 states provided some type of auditing procedures; with the most common being random site visits to districts and schools during test administration periods. The other 30 states provided no evidence that policy implementation was ever audited at the school or district level. State departments of education could improve test security by giving more attention to auditing the implementation of their policies.

States can also improve test security through statistical analyses of student answer sheets designed to detect manipulations. The average state policy earned 16% of the maximum possible points in the *statistical analyses* subcomponent. Only 14 states provided evidence that answer sheets or test scores were regularly screened through some type of statistical analysis. The level of sophistication of these analyses varied greatly

from state-to-state. The most common analysis, utilized by 11 states, involved looking for unusually high score gains at the school or district level. 5 states (Louisiana, Michigan, Mississippi, Texas, and Wisconsin) analyzed answer sheets for unusual patterns of erasures and 6 states (Illinois, Michigan, Mississippi, Pennsylvania, Texas, and Wisconsin) attempted to detect aberrant response patterns from examinees. Even though 14 states claimed to conduct statistical analyses, only 4 states (Illinois, Louisiana, Texas, and Wisconsin) provided access to reports of the results from these statistical analyses.

Table 4.5 Oversee subcomponent distributions

Oversee test preparation, administration, and scoring activities	Number of states earning 0	1	2	Avg.
• Test Security Audits	Mean: 0.62 (31% of max)			
Implementation of test security policy is audited regularly	30	9	11	0.62
• Statistical Analyses	Mean: 1.62 (16% of max)			
Provides for statistical analyses to detect possible manipulations	36	2	12	0.52
Score fluctuation analyses	39	1	10	0.42
Evidence of security analysis reports	41	5	4	0.26
Erasure analyses	45	0	5	0.20
Aberrant response analyses	44	1	5	0.22
• Score Reports	Mean: 0.22 (11% of max)			
Outlines procedures in making any changes to test scores	43	1	5	0.22
• Test administration oversight	Mean: 0.40 (10% of max)			
Provides for independent monitoring of test administration	36	11	3	0.34
Teachers are not to administer the test to their own students	47	3	0	0.06

The average policy earned 11% of the possible points in the *score reports* oversight subcomponent. This was because only 6 state policies addressed procedures educators can use when they suspect scores (or demographic information) might be incorrect. 43 states apparently either do not allow changes or leave it up for educators to decide appropriate actions when they suspect score information may be inaccurate.

The final oversight subcomponent that provides the greatest opportunity for improvement was in *test administration oversight*. The average state policy only earned 10% of the possible points in this subcomponent. This was due to the fact that only 3 states (Mississippi, Pennsylvania, and Texas) provide for independent monitoring of test administration. Another 11 states provide at least some level of test administration monitoring, with principals or proctors (who are school employees) witnessing test administration of at least a sample of classrooms. No state prohibits educators from administering tests to their own students, but 3 states (Mississippi, Nevada, Oklahoma) at least mentioned the potential for manipulations when educators administer tests to their own students.

With scores of 14 out of 18 possible points, Texas and Mississippi had the highest quality policies with regards to oversight. This was due, in large part, to both states implementing statistical analyses to identify potential test score manipulations. Mississippi, in particular, analyzes student answer sheets to check for unusual patterns of erasures, aberrant patterns of student responses, and unusually large score fluctuations from year to year. The state also allowed the results from one of these analyses to be published online (Caveon, 2008).

Mississippi's security policy also outscored other policies because of the way it requires test administrations to be monitored. In addition to school testing coordinators being required to monitor test administration procedures at randomly selected classrooms, the policy requires at least two trained individuals to be present from the time tests are distributed until the time in which test materials are returned to a secure area (Mississippi Department of Education, 2006). While this isn't as strong of a requirement as requiring test administration monitors to be completely independent, it should help to ensure that test scores are not manipulated during test administration.

Component Score: Inform

The third component to evaluate state test security policies determined to what degree the policies informed educators about why some behaviors and activities are unacceptable. Table 4.3 shows that of 44 points possible in this component, the median state policy earned 22.5 points (51%). Table 4.6 displays the subcomponent distributions for this component ranked by relative strength.

The subcomponent with the greatest relative strength was with regards to training, with the average policy earning 58% of the total possible points. Around 40 state policies provided for some type of training of district- and school-level testing personnel. More than half of all state policies required district test coordinators to be trained annually, while only 17 state policies required annual training for school coordinators. Only 16 state policies required annual training of all test administrators and proctors. 33 states provided training materials online, with 13 of those states having exemplary materials to train educators on test security.

The average state policy earned 52% of the possible number of points in *general guidance*. Within this subcomponent, the greatest relative strength was that around 40 policies specified that educators could not change student answers, give students answers, or provide hints to students. Some of the better policies even specified that educators could not use facial expressions or body language that may cause students to change their answers. 10 state policies did not specify that educators could not change student answers. 22 states provided detailed lists of materials that can or cannot be provided to students during testing and another 11 states at least identified some materials that cannot be provided to students. State policies were evenly split between those that did and did not specify that educators cannot read certain sections of the test aloud to students. The relative weaknesses of state policies in this subcomponent deal with test preparation and test sanitization guidance. 38 states did not limit the amount of time (or limit which activities) that can be used for test preparation. Of the 12 states that did limit test

preparation, some provided vague ethical guidelines and others provided specific test preparation time limits (and test preparation materials, such as practice tests). 33 state policies also did not provide guidance to educators as to whether or not they can sanitize answer sheets before tests are scored. 9 states did specify that educators either can or cannot clean erasures and stray marks on the answer sheets before scoring. Providing more detailed guidance in these areas may improve test security.

Table 4.6 Inform subcomponent distributions

Inform educators why some behaviors/activities are unacceptable	Number of states earning 0	1	2	Avg.
• Training	Mean: 5.80 (58% of max)			
Provides for regular training of district-level testing coordinators	9	15	26	1.34
Regularity / amount of training	11	12	27	1.32
Provides for regular training of school-level testing coordinators	10	23	17	1.14
Provides for regular training of all test proctors	12	22	16	1.08
Quality of training materials available online	17	20	13	0.92
• General Guidance	Mean: 6.20 (52% of max)			
Specifies that educators cannot give students hints or answers	7	4	39	1.64
Specifies that educators cannot change student answers	10	0	40	1.60
Specifies materials that can be provided to students	17	11	22	1.10
Specifies that educators cannot read certain sections aloud	21	9	20	0.98
Outlines procedures for retesting or sanitizing answer sheets	33	8	9	0.52
Limits the amount of time spent on test preparation activities	38	6	6	0.36
• Examples of appropriate and inappropriate behaviors	Mean: 4.56 (46% of max)			
Examples of accommodations	4	22	24	1.40
Appropriate = 22.0 Inappropriate = 4.2	--	--	--	--
Examples of test administration activities	4	22	24	1.40
Appropriate = 4.5 Inappropriate = 7.3	--	--	--	--
Examples of test preparation activities	11	30	9	0.96
Appropriate = 3.5 Inappropriate = 2.7	--	--	--	--
Examples of uses/interpretation of scores	27	17	6	0.58
Appropriate = 1.2 Inappropriate = 1.5	--	--	--	--
Examples of school/class activities on test day	39	11	0	0.22
Appropriate = 0.3 Inappropriate = 0.2	--	--	--	--
• Principles and Rules	Mean: 5.38 (45% of max)			
Explains that all students must be tested	14	8	28	1.28
Explains copyright laws/penalties for violating copyright	14	11	25	1.22
Explains the importance of standardized test administration	25	8	17	0.84
Refers to an honor code or code of ethics	30	4	16	0.72
Explains the uses of test scores	31	4	15	0.68
Explains the importance of validity & generalizing from test scores	30	8	12	0.64

State policies, on average, earned 46% of the total possible points in the *examples of appropriate and inappropriate behaviors* subcomponent. Within this component, state policies earned a score of 0 if the policies did not specify any examples. State policies that listed a total of 10 or fewer examples earned a score of 1; and policies with more than 10 examples (including at least one appropriate example and one inappropriate example) earned a score of 2. By far, the category with the greatest number of examples was the accommodations category. On average, state policies listed 22 appropriate accommodation examples and 4.2 inappropriate accommodation examples. Because many states had the same list of accommodation examples, the relative strength of state policies in this area might be due to assistance from the federal government or other organizations. In the other categories, state policies provided an average of 11.8 test administration activities (4.5 appropriate and 7.3 inappropriate); 6.2 test preparation activities (3.5 appropriate and 2.7 inappropriate); 2.7 score use examples (1.2 appropriate and 1.5 inappropriate); and 0.5 school or class activity examples (0.3 appropriate and 0.2 inappropriate). The state with the greatest number of examples was Michigan, with 135 specific examples of appropriate and inappropriate behaviors. Policies from Nebraska and Iowa provided no examples of specific testing behaviors. States could consolidate these lists (as appropriate to match their formalized beliefs regarding testing) to possibly improve test security.

The relatively weakest subcomponent was with *principles and rules*, with the average policy earning 45% of the total possible points. In this subcomponent, most state policies (all but 14 policies) specified that all students must be tested and that tests cannot be copied due to copyright laws. Exactly half of all policies explained the importance of standardized test administration procedures to deter manipulations. Only 20 state policies provided an honor code or code of ethics for educators to follow during test preparation, administration, and scoring activities. Seeing as how honor codes are highly

recommended by McCabe and Trevino (2002) to deter students from cheating, perhaps other states could develop honor codes to improve test security. The areas for most improvement in this subcomponent are explaining to educators how test scores are used (only 19 state policies provide this information) and explaining the importance of validity and making generalizations from test scores (only 12 state policies provide an explanation of this).

With a score of 41 out of 44 points possible, Michigan's security policy was clearly the best in informing educators about testing. It earned a high score in this component because of they way it clearly informed educators about the importance of test security, standardized testing procedures, and the role test security plays in helping to ensure valid inferences from test scores. The policy also clearly stated that educators cannot copy test materials, change student answers, give hints to students during testing, provide materials to students during testing, or sanitize student answer sheets after testing. In explaining the appropriateness of test preparation, administration, and scoring practices, the policy described 135 specific examples. After eliminating all the examples of accommodations for special education students, the policy still described 30 appropriate activities and 24 inappropriate activities. This included a statement explaining that educators should not, "... sacrifice significant instructional time by devoting large amounts of instructional time to commercially or locally prepared programs or drill-type assessment preparations" (Michigan Department of Education & Office of Educational Assessment and Accountability, 2007, p. 20). Michigan's policy was one of the few that attempted to limit test preparation activities.

The other reason why Michigan's policy scored so high in this component is due to its training provisions. The policy required all district testing coordinators, school testing coordinators, test administrators, and test proctors to be trained annually. While other state policies had these provisions, Michigan was unique in not only providing very detailed training materials (Michigan Department of Education & Office of Educational

Assessment and Accountability, 2007), but also providing educators with access to an assessment listserv to ask test security questions (Michigan Department of Education, 2008). These resources could help ensure educators clearly understand test security procedures and which behaviors are appropriate or inappropriate.

Component Score: Limit

The final component used to evaluate state test security policies determined to what degree the policies limit opportunities for educators to manipulate test scores. Table 4.3 shows that of 12 points possible in this component, the median state policy earned 6 points (50%). Table 4.7 displays the subcomponent distributions for this component ranked by relative strength.

Table 4.7 Limit subcomponent distributions

Limit opportunities for educators to manipulate test scores	Number of states earning 0	1	2	Avg.
• Materials Security	Mean: 5.64 (47% of max)			
Requires test materials to remain sealed until testing	8	2	40	1.64
Provides for a tracking system of test materials	10	10	30	1.40
Specifies who has access to test materials (and at what times	11	10	29	1.36
Specifies the amount of time materials are available	31	6	13	0.64
• Test Forms	Mean: 0.60 (15% of max)			
Requires new test forms (different test items) to be used annually	41	2	7	0.32
Requires multiple test forms (which may be reused)	43	0	7	0.28

Most state policies limit opportunities for educators to manipulate test scores by limiting access to testing materials. Around 40 state policies required test materials to remain sealed until testing, specified who has access to test materials, and provides a system to track all test materials through the entire testing process. Fewer than 20 state policies limit the amount of time test materials are available, with several states allowing educators (with permission) to examine test materials at any time.

The relative weakness in this area deals with the test forms used each year during testing. Only 7 state policies required new test forms (or test questions) to be used each year or multiple test forms to be re-used in alternating years. This might underestimate the actual number of states using new or multiple test forms each year, because this information might not be located within test security policy documents. Also, even though using new test forms each year would be an effective way to limit opportunities for manipulations, this is an extreme (and costly) measure. Thus, even though state policies scored poorly in this subcomponent, it is not recommended that all states require new test forms annually.

With a perfect score of 12 out of 12, the policy from Illinois outscored all other policies in limiting opportunities for educators to manipulate test scores. This was because the Illinois policy was strongest ensuring the security of testing materials. Not only does the policy require materials to remain sealed until testing, it also specifies who has access to the materials (and for how long materials can be accessed) and provides for tracking of all materials. The primary way in which the Illinois policy outscored other policies in this component is by its requirement that new test forms be used each year. This helps limit the ability of educators to manipulate test scores through knowledge of test items.

Relationship Among Subcomponent Scores

To investigate relationships among the subcomponent scores, all 105 possible pairs of Spearman rank-order correlations were examined. Of these 105 correlations, 100 were positive (and 5 were statistically significant at a Bonferroni-adjusted level of 0.05). 21 correlations were found to be greater than 0.50, with the biggest correlation of 0.7537 found between the training and implementation subcomponents. This makes sense, because states with better training would be expected to have strong implementation

scores as well. Of the 5 negative correlations, none were found to be statistically significant.

State Policy: Dichotomizations

Table 4.3 displays the number of state policies categorized within each of seven dichotomizations. While this information will be used in an analysis later in this chapter, Table 4.8 displays which state policies fell into which of the categories. The table shows that due to some ambiguity in state test policies, some additional categories had to be created (such as for policies that are implemented equally at the state- and district-levels).

Table 4.9 displays the mean composite evaluation scores for states falling in each category. For example, states with clear and accessible policies earned an average composite score of 64.9 compared to an average score of 39.7 for states with ambiguous or difficult-to-find policies. Likewise, states with policies representing state-level mandates (with state-level laws, rules, and regulations) outscored states that left most of the policy work to individual school districts. States with policies that provide for independent monitoring of test administration outscored states with no independent monitoring, and state policies with many specific examples outscored, on average, policies with fewer examples.

Table 4.9 displays the mean composite evaluation scores for states falling in each category. For example, states with clear and accessible policies earned an average composite score of 64.9 compared to an average score of 39.7 for states with ambiguous or difficult-to-find policies. Likewise, states with policies representing state-level mandates (with state-level laws, rules, and regulations) outscored states that left most of the policy work to individual school districts. States with policies that provide for independent monitoring of test administration outscored states with no independent monitoring, and state policies with many specific examples outscored, on average, policies with fewer examples.

Table 4.8 Policy Categorizations

Clear and accessible	Ambiguous/hard-to-find		Unclassified
AK, AZ, AR, CT, DE, FL, GA, HI, ID, IL, KS, KY, LA, MD, MI, MS, MT, NV, NC, ND, OK, PA, SC, SD, TX, UT, VA, WA, WV, WI	AL, CA, CO, IN, ME, MA, MN, MO, NE, NH, NJ, NM, NY, OH, OR, RI, TN, VT, WY		IA
State-level	District-level	Both levels equal	Unclassified
AZ, CA, CT, DE, HI, ID, KY, MA, MT, NM, NY, NC, OK, PA, SC, TX, UT, VT, VA, WA, WI, WY	AL, AK, AR, CO, GA, IL, IN, IA, KS, ME, MD, MO, NE, NJ, ND, OH, OR, SD, TN	LA, MI, MN, MS, NV, WV	FL, NH, RI
Punitive or law-focused	Instructive/informative		Unclassified
AL, CA, CT, DE, KY, LA, MD, MA, MS, NV, NY, OH, OK, TN, TX, WV	AK, AZ, CO, FL, GA, HI, ID, IL, IN, MI, MN, MT, NM, NC, ND, OR, PA, SC, SD, UT, VA, WA, WI		AR, IA, KS, ME, MO, NE, NH, NJ, RI, VT, WY
Independent monitoring	No independent monitoring		Unclassified
AL, IL, KY, LA, MI, MS, OK, PA, TN, TX, WA	AK, AZ, AR, CA, CO, CT, DE, FL, GA, HI, ID, IN, IA, KS, ME, MD, MA, MN, MO, MT, NE, NV, NH, NJ, NM, NY, NC, ND, OH, OR, RI, SC, SD, UT, VT, VA, WV, WI, WY		
Investigative	Preventative	Equal	Unclassified
CT, DE, GA, KY, LA, MD, MA, MN, MS, MT, NV, OH, PA, SC, TN, TX, WV	AL, AK, AZ, AR, CA, CO, FL, HI, ID, IN, KS, MI, NJ, NM, NY, NC, ND, OK, OR, SD, UT, VA, WA, WI	IL	IA, ME, MO, NE, NH, RI, VT, WY
Example-based	Not many examples		Unclassified
AL, AK, AZ, AR, CA, CO, GA, IL, KS, KY, MA, MI, MN, NM, NC, OR, TX, UT, VA, WA, WI	CT, DE, FL, HI, ID, IN, IA, LA, ME, MD, MS, MO, NE, NV, NH, NJ, NY, ND, OH, OK, PA, RI, SC, SD, TN, VT, WV, WY		MT
Positive message	Negative message	Both	Unclassified
AK, AZ, CO, FL, GA, HI, IN, KS, MI, MN, MT, NM, NC, ND, OR, SD, UT, VA, WA, WI	AL, AR, CA, CT, DE, KY, LA, MA, MS, NV, NY, OH, OK, PA, SC, TN, TX, WV	IL, MD	ID, IA, ME, MO, NE, NH, NJ, RI, VT, WY

Other categorizations of policies did not show large differences in average composite scores. For example, policies that focused on instructing and informing educators about test security only earned a slightly higher average composite score than

policies that focused on sanctions and legal reasons for test security. Likewise, policies that stressed procedures for investigating potential manipulations only scored slightly higher, on average, than policies that focused on preventing manipulations. Finally, policies that had an overall positive tone regarding test security only slightly outscored policies with a negative tone.

Table 4.9 Policy evaluation composite scores by category

Group:	Policies	Mean Composite	Std. Deviation	p-value
Clear and accessible	30	**64.9**	17.8	<.0001*
Ambiguous or difficult-to-find	19	39.7	17.9	
State-level mandate	22	**62.5**	17.9	.0023*
District- or school-level responsibility	18	40.6	20.7	
State and district responsibility (equally shared)	6	74.0	15.0	--
Punitive or law-focused	16	59.4	18.4	.3788
Instructive/informative	23	64.4	16.1	
Independent monitoring	11	**72.5**	19.7	.0015*
No independent monitoring	39	48.9	20.7	
Investigative	17	63.7	16.5	.6300
Preventative	23	56.5	18.8	
Equally investigative and preventative	1	89.0	--	--
Example-based	21	**65.0**	20.3	.0028*
Not many examples	28	46.0	21.4	
Positive message	20	62.1	16.5	.9951
Negative message	17	60.9	17.8	
Combination of positive and negative message	2	64.0	35.4	--

Other Categories of Security Policies

Examining the policy evaluation data for each state, five other policy categorizations emerged. These categories include states with reactionary policies, states with exceptional training methods, states with policies that limit test preparation activities, states that focus on the financial aspects of test security, and states with no formal policies. A brief description of the policies in each of these categorizations provides more insight into how states attempt to limit test score manipulations.

Reactionary policies

Reactionary policies are those that are developed or modified in reaction to specific security breaches. While modifying a policy after a security breach shows a state is continually working to improve its policy, it also demonstrates that the original policy was inadequate in preventing the security breach. Six states (Hawaii, Illinois, Kentucky, Massachusetts, Texas, and Washington) provide evidence of reactionary test security policies.

Some states made minor changes to their policies in response to security breaches. In 2007, Hawaii increased test security after an unspecified test security incident the previous year. The reaction to this incident was that state superintendent Hamamotohas required all principals, vice principals, and test coordinators to sign an attendance log recording their attendance at a mandatory training session. The training session informed educators how to implement "heightened test security procedures at their schools" (Hawaii Department of Education, 2007, p. 2). Likewise, Illinois made small changes to its security policy following a security breach in 2005. The breach was described in a letter from the state superintendent, Randy Dunn:

> The Illinois State Board of Education (ISBE) has been investigating an ISAT security breach with statewide implications for the 2005 ISAT administration. The breach involved a grade 5 reading passage and the multiple-choice and extended-response questions about the passage that were exposed prior to testing. ISBE was notified about the breach on Tuesday, March 8, 2005, one day after the start of the ISAT testing window. ISBE test contractors have been advised not to score the affected questions and, for grade 5 students statewide, to use only students' performance on the remaining test questions to generate grade 5 ISAT reading scores. Our test contractors have assured us that the deletion of the items associated with this passage will not have an effect on reliability and validity of the grade 5 reading test results for schools. All grade 5 reading results will be reported as usual for students, school districts and the state (STATESUP, 2005, p. 3)

Following this incident, Illinois published an updated version of its *Professional Testing Practices for Educators* (Illinois State Board of Education, 2006) strengthening procedures to ensure the security of test materials.

Officials in other states made more drastic modifications to their test security policies in response to specific incidents. In a 2005 memo from the Superintendent of Public Instruction entitled, *Critical Updates to Assessment Security*, the following explanation for increased test security was made:

> Recent events make clear that the Office of Superintendent of Public Instruction (OSPI) must implement several measures to maintain and improve testing security; the measures are critical to maintaining the validity of our assessment system. This $325,005 request is designed to mitigate opportunities for in advertent viewing of test materials prior to administration of the assessment, provide parent access to their child's assessment in a secure and appropriate manner, and establish protocol to define professional practices related to test administration. The request is a direct result of several recent events that may compromise our assessment integrity in the near future; events unanticipated for adoption in the biennial operating budget. (Washington Office of Superintendent of Public Instruction, 2005, p. 1)

Before this memo, reading and writing test booklets were combined for grades 4-7. The memo made the case that separating the test materials for each grade was. "...necessary for continued test security in order to mitigate the occurrence of inadvertent viewing," (p. 1) and that "Separating the books will prevent a student or proctor from viewing a second content area's material while administering another" (p. 1).

The policy in Texas is another one that has evolved as a result of specific test security breaches. Table 4.10 displays a sample of news reports and memos from the Texas Education Agency (TEA) outlining how security incidents and evaluations influenced decisions to modify the state test security policy between 2005 and 2007. The table shows how reports of specific incidents led to evaluations of the security policy by testing experts and educator-led committees. These evaluations led to lists of recommendations, including the hiring of an Inspector General and the implementation of statistical analyses. These statistical analyses flagged large numbers of potential score manipulators, which led to further additions to the security policy. Even when many of the educators flagged by the statistical analyses were found to be innocent, the additions to the security policy remained. The story of the reactionary nature of the Texas state test

security policy slows with the resignation of the Commissioner of the Texas Education Agency on July 1, 2007. Ironically, the news report of this resignation states that the TEA had been "... under fire for what some have seen as a slow, halfhearted response to questions about how frequently students and educators are cheating on the Texas Assessment of Knowledge and Skills" (Embrey, 2007, p. 1).

Massachusetts is another state with a seemingly reactionary policy. In 2000, after receiving complaints from teachers, the Commissioner of Education sent a memo to principals stating that teachers would no longer be required to sign an agreement about test security (Driscoll, 2000). This seems to be the last time security procedures were relaxed. In 2007, the Commissioner reported that almost 4000 students who did not have IEPs received accommodations during testing (Driscoll, 2007) and that a "larger-than-expected percentage of students with disabilities received one or more nonstandard accommodations" (p. 1). Perhaps because of these findings, the Massachusetts test security policy provides 30 extremely specific examples of testing irregularities. It seems as though these examples are simply summaries of problems encountered during previous administrations of the state test.

Another state that reacted to a specific incident by modifying its security policy is Kentucky. In test administration training materials from 2006, the state received the following security breach allegations: test security (732 reports), out-of-order testing (65 reports), inappropriate assistance by staff (43 reports), special education irregularities (36 reports), missing test booklets (7 reports), and other reports (32 reports) (Kentucky Department of Education, 2006). In response to these reports, the state created additional training modules incorporating case studies and quizzes for educators to test their understanding of test security.

Table 4.10 Test security incidents and responses in Texas from 2005-07

Document Title	Date	Description
Students may have been helped. (Garza, 2005)	05/03/2005	Published report of security incident. See Appendix A.
Houston Chronicle Report (CEA, 2007)	05/05/2005	Published report of security incident. See Appendix A.
Texas has zero tolerance for those who cheat students by cheating on test (TEA, 2005)	01/10/2005	The TEA hires a testing expert to review test security measures, reminds educators of the consequences of manipulating test scores, and builds a tracking system to help monitor test irregularities. The document then describes the state test security policy.
Caveon Data Forensics Pilot Report (Maynes, 2005)	Spring 2005	Test security company Caveon analyses data from more than 5 million tests from the 2005 administration of the TAKS in Texas. The report finds statistical anomalies in 1% of classrooms and 8.6% of schools. Caveon recommends the state develop an investigative process for anomalous findings.
Final Report: Review and Recommendations Related to Test Security (Cizek, 2005)	07/27/2005	Testing expert, Gregory Cizek, completes his evaluation of test security in Texas. He lists 11 recommendations to improve security. While test materials security was a strength of the current policy, Cizek recommends the state improve the procedures used to report test irregularities.
Analysis suggests cheating on TAKS. (Benton, 2006a)	05/23/2006	Statistical report from Caveon. See Appendix A for summary
Big Spring's TAKS tests flagged. (Levesque, 2006)	06/09/2006	Published report of security incident. See Appendix A.
Inquiry targets 20 area schools. (Brock, 2006)	07/28/2006	The TEA investigates testing irregularities at 609 schools. See Appendix A.
2 HSID teachers resign in test-cheating probe. (Tresaugue & Viren, 2006)	07/30/2006	Published report of security incident. See Appendix A.
Commissioner names task force on test security. (TEA, 2006a)	08/01/2006	As a result of the 2005 analysis from Caveon, the state names a task force on test security to examine security issues and oversee investigations of campuses and districts with alleged testing anomalies.
Cheating hasn't hurt Wilmer-Hutchins teachers. (Benton, 2006b)	10/01/2006	News report noting that teachers found guilty of test score manipulations are now working in other public schools in Texas. See Appendix A for summary.
Inspector General hired to oversee test irregularity investigations. (TEA, 2006b)	10/05/2006	Based on recommendations from the task force, an Inspector General was hired to oversee reports of testing irregularities. Educators will be allowed to submit reports of suspected manipulations anonymously. The Inspector General will be given subpoena power and will conduct audits into test security procedures at school districts.
More than 590 schools cleared of testing irregularities. (TEA, 2006c)	12/14/2006	Investigations into possible testing irregularities at 590 Texas schools are closed because no evidence of wrongdoing was found.

Table 4.10 Continued

Task force on test integrity recommends security enhancements. (TEA, 2007d)	01/04/2007	The task force offers 10 recommendations to improve test security, including the use of statistical analyses and improvements to more quickly investigate potential manipulations.
Commissioner's response to recommendations from the task force on test integrity. (TEA, 2007f)	Winter 2007	The TEA accepts all 10 recommendations from the task force on test integrity. Three of the recommendations were accepted with modifications.
Testing audits closed at 88 schools. (TEA, 2007c)	02/28/2007	Audits into test security procedures and possible testing irregularities at 88 schools are being closed after no evidence of improprieties was uncovered.
Winona ISD had possible TAKS security breach. (Waters, 2007)	05/01/2007	Published report of security incident. See Appendix A.
Test security enhancements planned. (TEA, 2007e)	05/31/2007	The TEA reports that statistical analyses for unusual score fluctuations, erasure patterns, or large number of students absent will be implemented during the 2006-07 school year.
Estimated number of cheaters might be low. (Benton & Hacker, 2007a, 2007b)	06/04/2007	A statistical analysis conducted by a professor at McMaster University in Canada finds that the scores from more than 50,000 students show evidence of cheating that could include students copying answers from other students or educators doctoring student answer sheets. See Appendix A.
Sanctions recommended against three schools and three educators because of testing improprieties. (Texas Education Agency, 2007a)	06/14/2007	Published report of security incident. See Appendix A.
FW charter school in trouble over TAKS cheating. (Benton, 2007a)	06/15/2007	Published report of security incident. See Appendix A.
Allegations of TAKS cheating. (McCollum, 2007)	06/21/2007	Published report of security incident. See Appendix A.
Texas' education commissioner to resign. (Embrey, 2007)	07/01/2007	State Education Commissioner Shirley Neeley resigns after serving as Commissioner of the TEA for 3.5 years. The report describes the TEA as having been "... under fire for what some have seen as a slow, halfhearted response to questions about how frequently students and educators are cheating on the Texas Assessment of Knowledge and Skills" (p. 1).
TEA: teacher leaked parts of TAKS test. (Benton, 2007b)	07/13/2007	Published report of security incident. See Appendix A.

Policies with exemplary training procedures

As just mentioned, the state of Kentucky developed online training modules to help educators understand the importance of test security and choose appropriate activities during testing. In addition to a self-administered quiz for educators to test their knowledge of which activities are appropriate or inappropriate, the Kentucky Department of Education developed a story of the fictitious Mr. Reddy as he prepares, administers, and returns materials from an administration of the state test (Kentucky Department of Education, 2006). This entertaining approach to training educators about test security might be recommended over the more traditional training materials found in other states.

Washington and Pennsylvania take a similar approach to Kentucky in providing questionnaires and self-tests with their test security training materials. Whereas Washington's self-test simply allows educators to check their level of understanding, Pennsylvania's questionnaires allow educators to also evaluate the quality of the training they receive.

While many states provided online training materials, Kansas requires a face-to-face approach to training. Before the 2006-07 test administration, the Kansas Department of Education provided a two-day assessment conference for test coordinators, school administrators, and teachers. The state of Oklahoma also holds annual in-service training on test security each year. Oklahoma differs in that it requires school districts to send a minimum of one representative from each school to these in-services.

States that limit test preparation activities

The 2007 training in-service offered by the Oklahoma Department of Education focused on test preparation practices. Likewise, training materials from Wisconsin also focus heavily on test preparation activities. These states represent a third group of policies – policies that focus on limiting educators' test preparation activities.

While many states make little mention of which test preparation activities are appropriate or inappropriate, the policy from Colorado requires educators to receive approval for any materials used in preparation for testing. Indiana's policy goes one step further in stating that, "any activity in the school or classroom, overt or inadvertent, that creates an excessive focus on the specific test content of ISTEP+, for the purpose of artificially raising test scores, is inappropriate" (511 IAC 6.1-1-4).

The states of New Hampshire, Rhode Island, and Vermont take a different approach to deter manipulations via test preparation. Rather than limiting what educators are allowed to do in preparing for the test, these states provide educators with access to state-sanctioned practice tests. Perhaps the belief is that if you give educators practice tests, they will be less likely to develop their own practice tests with knowledge gained from previous test administrations.

States focusing on financial aspects of security

The next categorization of test security policies includes those that have a financial aspect. For example, the policy in Delaware clearly states that educators found to have violated the policy will face civil sanctions and must pay any costs incurred by the State as a result of the violation (Delaware Code, 2001). The policy in Florida, which allows for a $1000 fine for violations, also uses financial sanctions to deter educators from manipulating test scores (Florida Department of Education, 1999).

In addition to providing financial sanctions for those found violating its policy, the state of Minnesota makes a financial argument for the importance of test security. In the *Procedures Manual for the Minnesota Assessments* (Minnesota Department of Education, 2007), the argument is made that, "Test security is needed to preserve the integrity of the tests and test results and to protect the state's financial investment from compromise" (p. 25). Perhaps financial incentives are more compelling to educators than

the traditional arguments that test security is important in order to maintain the validity of inferences made from test scores.

States with no policies

The final categorization of state test security policy includes states with no formal policies, at least at the state-level. The most obvious example of this comes from the state of Iowa. In Iowa, the only state policy regarding test security is found in the Iowa Administrative Code 292 – rule 25.(3)e, which states:

> Falsifying or deliberately misrepresenting or omitting material information regarding the evaluation of students or personnel, including improper administration of any standardized tests, including, but not limited to, changing test answers, providing test answers, copying or teaching identified test items, or using inappropriate accommodations or modifications for such tests" (Iowa Board of Educational Examiners, 2004).

The state has no other formal security policy statement. Following a reported incident of an educator manipulating test scores in 2005, the Iowa Department of Education and the Iowa Testing Programs issued letters to school districts in Iowa encouraging them to develop test security policies. While districts were also given a sample security policy, they have never been required to adopt or implement specific test security measures. For a more detailed description of the policy in Iowa, see Thiessen (2007).

Other states also leave most of the test security work to individual school districts. Policies from Maine, Missouri, Nebraska, and New Jersey all state that test security is the responsibility of school districts. These states differ from Iowa in that they provide guidelines and requirements for these district-developed policies. Thus, even though these states do not have a single state policy, the individual district policies must meet at least a basic level of quality.

Changes in Policies Over Time

As discussed earlier, any changes in state test security policy content or procedures from 2005-07 were used to conduct longitudinal analyses of the relationship between policy quality and score trend discrepancies. Table 4.11 shows that 16 states updated or modified policy content or procedures during this period. The table also shows the types of modifications made to the policies during this time.

The table shows that some states developed additional policy documents from 2005-07. Alabama published an ethical code for testing in 2005 and Arkansas first published all its policy content online in 2007. Pennsylvania also first published a list of examples of appropriate and inappropriate testing activities in 2006. Instead of publishing a new document, Oregon updated its training materials in an attempt to improve test security procedures.

Table 4.11 Changes in state policy content or procedures from 2005-07

State	Year	Modifications	Excluded?
Alabama	2005	Published ethical code	No 2003-05
Arkansas	2007	Published policy content online	No 2003-05
California	2005	Unknown	Unknown
Georgia	2005/07	Developed security policy laws	No data
Hawaii	2007	Heightened security following incident	--
Illinois	2007	Response to 2005 incident (materials)	No 2005-07
Kentucky	2006	Significant changes after '05 evaluation	No 2003-05
Louisiana	2005	Added erasure & fluctuations analyses	--
Michigan	2005	Minimal changes to ethical practices	Insignificant
Mississippi	2006	Added all statistical analyses	--
Ohio	2007	Unknown	Unknown
Oregon	2007	Modified training materials	No data
Pennsylvania	2006	Published "Do's and Don'ts" document	--
Texas	2005/06/07	Responses to evaluation & incidents	No data
Washington	2005	Added security budget after incident	--
West Virginia	2006	Updated investigation procedures	No 2003-05

Other states made more significant changes. Georgia first added test security rules to its state administrative code in 2005 and 2007. Two states, Louisiana and

Mississippi, implemented statistical analyses of answer sheets to their security procedures beginning in 2005 and 2006. Six states, Hawaii, Illinois, Kentucky, Texas, Washington, and West Virginia all modified the security policy content or procedures in reaction to security breaches. Washington, in particular, added a test security budget in response to a breach in the security of testing materials.

While these changes should provide an opportunity to analyze any changes in the relationship between policy quality and score trend discrepancies, a combination of two factors stand in the way of this analysis. First, the extent to which some states modified their policies is unknown. For example, policy documents indicate that California, Michigan, and Ohio modified policy content, but it is difficult to find what modifications were made. Second, several states cannot be included in a longitudinal analysis due to a lack of score trend discrepancy data. Alabama, Arkansas, Georgia, Illinois, Kentucky, Oregon, Texas, and West Virginia are all missing at least some data from 2003-07. Thus, longitudinal changes from 2003-05 to 2005-07 cannot be determined for these states.

The only states that made identifiable policy changes from 2005-07 and have data available to estimate score trend discrepancies from 2003-05 and 2005-07 are Hawaii, Louisiana, Mississippi, Pennsylvania, and Washington. With such a small sample size, any conclusions drawn from such an analysis should not be generalized to other states.

Summary: Test Security Policy Quality

The analyses in this section indicate state policies varied in quality both among and within the components of formalization, oversight, information, and limitations. Relative strengths and weaknesses of all states and comments about policies from specific states provide opportunities for states to improve test security procedures.

State-NAEP Trend Discrepancy Estimates

A scale-invariant method based on P-P plots was used to estimate the discrepancies between score trends on state tests and NAEP in reading and mathematics

for 4th and 8th grades from 2003-05, 2005-07, and 2003-07. First, results for each time period are considered; then results are considered separately for each grade-and-subject combination. Finally, results are discussed from analyses designed to evaluate some of the potential limitations of the V statistic.

Discrepancy Estimates for 2003-05, 2005-07, and 2003-07

Table 4.12 displays the discrepancies between scale-invariant trend effect sizes (V) for each state over each time period, grade, and subject. The first four columns of the table list the state along with the average state-NAEP trend discrepancy estimate for that state over 2003-05, 2005-07, and 2003-07. The last four columns of the table display the average two-year discrepancy separately for reading, mathematics, 4th grade, and 8th grade (averaging the 2003-05 and 2005-07 discrepancy estimates).

131 of the 183 (72%) average discrepancy estimates in the table are positive values, representing cases in which state trends were more positive than NAEP trends. Looking at the second-to-last row in the table, the median state experienced a state-NAEP trend discrepancy of 0.106 standard deviation units from 2003-05; .031 standard deviation units from 2005-07; and 0.070 standard deviation units from 2003-07. The last row shows that state trends were, on average, 0.084, 0.040, and 0.119 standard deviation units more positive than NAEP trends for 2003-05, 2005-07, and 2003-07, respectively. The last two rows also show that, on average, state trends were more positive than NAEP trends in both reading and mathematics, and for both 4th and 8th grades.

These results are better displayed in the scatterplots of Figure 4.3. The scatterplots display the relationship between state and NAEP trend estimates for each state-subject-grade combination. A star denotes the centroid of each scatterplot and the grids display both the number of observations falling in each quadrant and the number of states falling above and below the diagonal.

Table 4.12 Average state-NAEP discrepancy estimates

	2003-05	2005-07	2003-07	Average Discrepancies (2003-05, 2005-07)			
				Reading	Math	4th Grade	8th Grade
Alabama		-0.006		-0.059	0.048	-0.076	0.065
Alaska							
Arizona		-0.042		-0.076	-0.008	-0.031	-0.053
Arkansas		0.242		0.178	0.306	0.255	0.229
California	0.167	0.061	0.232	0.117	0.108	0.103	0.135
Colorado	0.096	-0.048	0.004	-0.029	0.070	0.026	0.004
Connecticut	-0.023	0.053	0.029	-0.005	0.044	<0.001	0.039
Delaware	0.119	0.181		0.243	0.057		0.150
Florida	-0.040	0.057	0.012	0.001	0.015	0.029	-0.012
Georgia							
Hawaii	0.044	-0.041		0.067	-0.092	-0.012	-0.013
Idaho	0.239	-0.147		-0.015	0.106	-0.064	0.156
Illinois	0.165			0.307	0.023		0.165
Indiana							
Iowa							
Kansas	0.214	-0.034	0.176	0.187	-0.007	-0.007	0.187
Kentucky		0.148		0.191	0.105	0.191	0.105
Louisiana	0.008	0.085	0.096	0.089	0.004	0.045	0.048
Maine	0.128	0.417	0.497	0.243	0.302	0.284	0.262
Maryland							
Massachusetts	-0.155	-0.038	-0.190	-0.127	-0.086	-0.118	-0.052
Michigan	0.184	0.154	0.335	0.189	0.159	0.196	0.116
Minnesota							
Mississippi	0.065	-0.006	0.062	0.059	0.001	0.036	0.059
Missouri							
Montana	0.030	0.091		0.075	0.045	0.048	0.073
Nebraska							
Nevada							
New Hampshire							
New Jersey							
New Mexico							
New York	0.115			0.107	0.123	0.106	0.124
North Carolina	0.035			0.098	-0.027	0.015	0.055
North Dakota		0.043		0.147	-0.060	0.039	0.047
Ohio		0.019		0.019		0.073	-0.036
Oklahoma	0.149	0.082	0.356	0.081	0.128	-0.022	0.168
Oregon							
Pennsylvania	0.151	0.106	0.281	0.143	0.114		0.128
Rhode Island							
South Carolina	0.074	-0.003	0.070	0.097	-0.027	0.068	0.003
South Dakota							
Tennessee							
Texas	0.179			0.168	0.190	0.148	0.210
Utah							
Vermont							
Virginia							
Washington	0.157	-0.154	0.004	0.051	-0.048	0.001	
West Virginia		-0.040		-0.034	-0.046	-0.046	-0.034
Wisconsin	-0.016	-0.013	-0.027	0.014	-0.043	-0.052	0.023
Wyoming	0.024			-0.015	0.064	0.004	0.044
Median State	**0.106**	**0.031**	**0.070**	**0.081**	**0.045**	**0.029**	**0.062**
Avg Observation	**0.084**	**0.040**	**0.119**	**0.089**	**0.058**	**0.053**	**0.094**

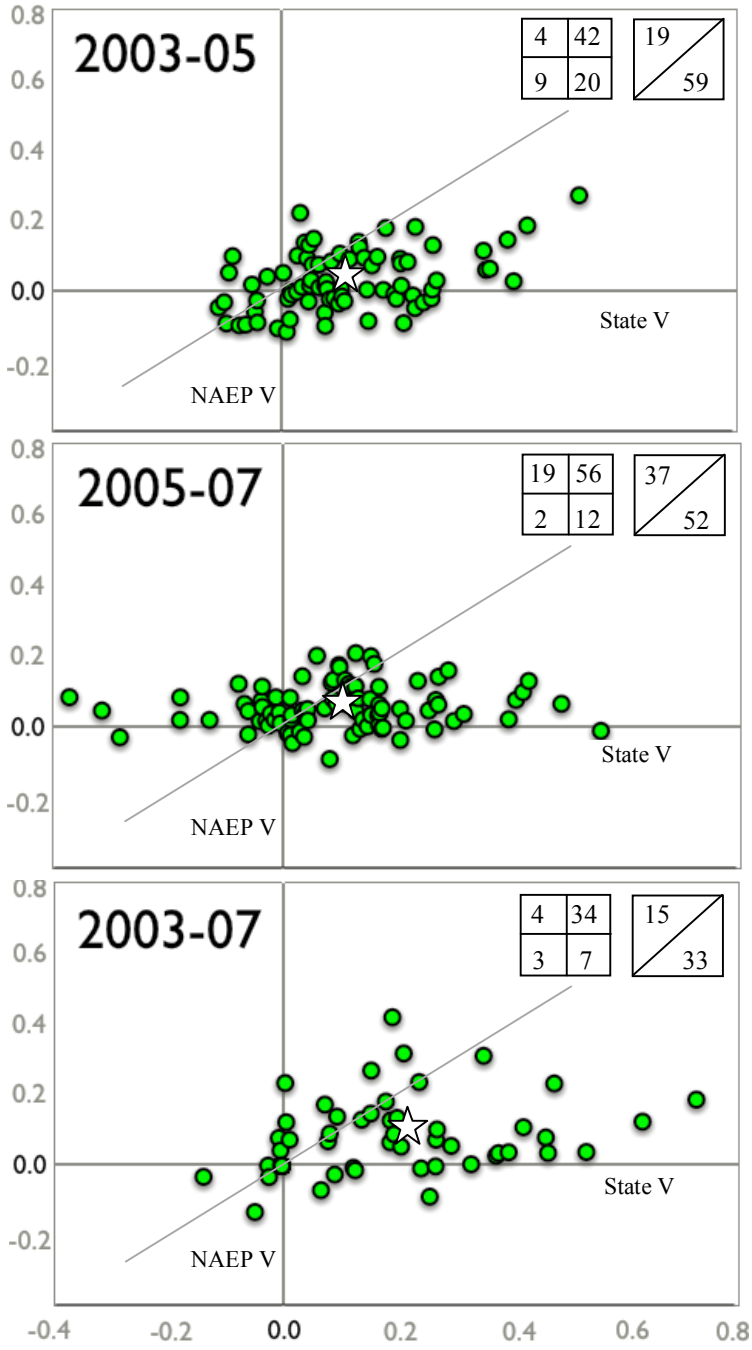
Note: Blank cells represent observations excluded from the study

The figure shows that for the 78 observations from 2003-05, the average trend effect size for state tests was estimated to be 0.118 standard deviation units. Over that same time period, the average NAEP trend was found to be only 0.034 standard deviation units. The average discrepancy between state and NAEP trends was, therefore, calculated to be 0.084 units. Thus, trends reported from states over this time period were 3.462 times higher than trends reported from NAEP.

The grid on the top-right of the first scatterplot in Figure 4.3 also shows the observations by quadrant and by their location with respect to the diagonal. For 2003-05, more than three-fourths, or 59 out of 78 (76%), observations were found to be below the diagonal. These are cases in which state trends were more positive than NAEP trends. 4 cases (5%), located in the first quadrant, showed positive NAEP trends and negative state trends, and 20 cases (26%) in the fourth quadrant showed positive state trends and negative NAEP trends. Thus, almost one-third of all cases showed a reversal in sign between state and NAEP trends.

These results all indicate that reported state trends were significantly higher than reported NAEP trends from 2003-05. A matched-pairs t-test confirms that the average state trend was significantly higher than the average NAEP trend for this period ($t_{78} = 5.938; p < .001$). Similar t-tests found statistically significant discrepancies for the 2005-07 ($t_{88} = 2.2954; p < .01$) and 2003-07 ($t_{47} = 4.2151; p < .001$) time periods.

While state trends were significantly more positive than NAEP trends for all three time periods, the discrepancies were smaller for 2005-07 than they were for 2003-05 trends. In fact, the average discrepancy was twice as large from 2003-05 than it was from 2005-07 (when state trends were 1.647 times higher than NAEP trends). Likewise, while 76% of observations reported higher state trends than NAEP trends for 2003-05, this fell to 58% of observations for 2005-07. The 2005-07 trends did show a greater percentage of observations with reversals in signs for state and NAEP trends, with 35% of observations falling in the first or fourth quadrants.



	Number of paired trends	Centroid (State, NAEP)	State-NAEP Discrepancy	State / NAEP	% of cases with State > NAEP
2003-05	78	(.118, .034)	.084	3.462	76%
2005-07	89	(.102, .062)	.040	1.647	58%
2003-07	48	(.210, .090)	.119	2.320	69%

Note: All discrepancies were found to be significant at $p<0.01$

Figure 4.3 State-NAEP discrepancies for 2003-05, 2005-07, 2003-07

The first thought after witnessing this decline in state-NAEP discrepancies from 2003-05 to 2005-07 was that, perhaps, states with large trend discrepancies in 2003-05 experienced small (or negative) discrepancies in 2005-07. If this were the case, explaining discrepancies for any given time period would be difficult (as these trends could simply just reverse over the next time period). To investigate this, the relationship between 2003-05 and 2005-07 discrepancies were plotted on the scatterplot in Figure 4.4.

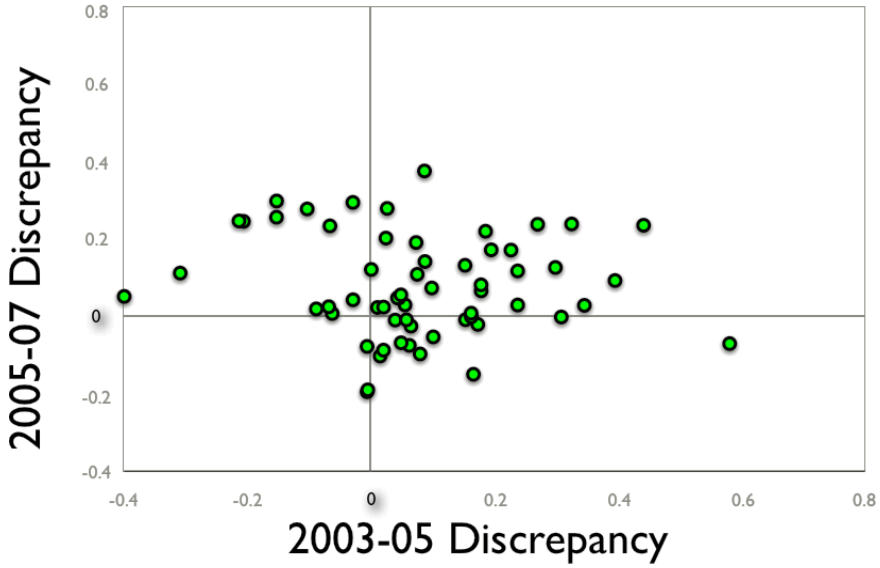


Figure 4.4 Relationship between 2003-05 and 2005-07 State-NAEP trend discrepancies.

Of the 60 cases from which both 2003-05 and 2005-07 trend discrepancies could be estimated, 31 (52%) of the cases experienced larger discrepancies in 2003-05 and 29 (48%) experienced larger discrepancies in 2005-07. The correlation between discrepancy estimates in Figure 4.4 was found to be -0.14, indicating a slight negative linear relationship. 5 cases (8%) reported NAEP trends more positive than state trends in both 2003-05 and 2005-07; 11 cases (18%) reported more positive NAEP trends in 2003-05 and more positive state trends in 2005-07; 18 cases (30%) reported more positive state

trends in 2003-05 and more positive NAEP trends in 2005-07; and 26 cases (43%) reported state trends more positive than NAEP trends in both time periods.

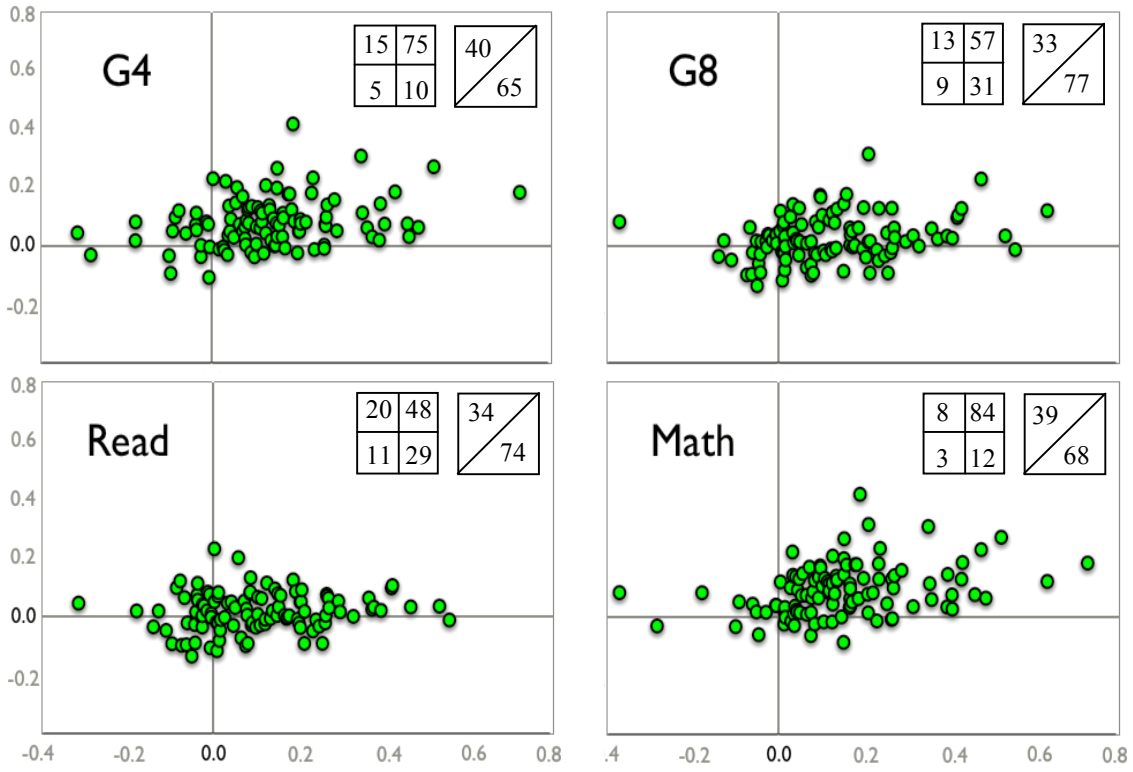
Over the four-year period from 2003-07, the results in Figure 4.3 show that state trends were 2.32 times larger than NAEP trends, with 69% of observations having state trends more positive than NAEP trends. So even though discrepancies were bigger for 2003-05 than for 2005-07, state trends were consistently larger than NAEP trends.

Discrepancy Estimates by Grade and Subject

To investigate if discrepancies differed by grade or subject, the scatterplots and table in Figure 4.5 were constructed. The results show that while state trends were significantly higher than NAEP trends for both 4th and 8th grade and for both reading and mathematics, the size of the discrepancies did differ for various subject-grade-year combinations.

The results show that discrepancies were larger for 8th grade (state trends 3.976 times larger than NAEP trends) than for 4th grade (state trends 1.611 times larger than NAEP trends). Discrepancies were also larger for reading (state trends 5.221 times larger than NAEP trends) than for mathematics (state trends 1.609 times greater than NAEP trends). The larger discrepancies in reading were expected, due to content differences. Some states only report English Language Arts (ELA) test results; these ELA tests would be expected to differ somewhat in content from the NAEP reading test.

Looking at subject-grade combinations, 8th grade reading trends showed the greatest average discrepancy. While states reported an average 8th grade reading trend of +0.112 standard deviation units, the average NAEP trend for these same states was -0.005 standard deviation units. The smallest state-NAEP discrepancy was found for 4th grade math, with an average state trend of +0.171 standard deviation units and an average NAEP trend of +0.125 standard deviation units.



	# of paired trends	Centroid (State, NAEP)	Discrepancy	State / NAEP	% of states with State > NAEP
4th Grade	105	(.139, .086)	.053	1.611	62%
8th Grade	110	(.125, .031)	.094	3.976	70%
Reading	108	(.110, .021)	.089	5.221	69%
Mathematics	107	(.154, .096)	.058	1.609	64%
4th Grade Read	53	(.108, .048)	.059	2.230	66%
4th Grade Math	52	(.171, .125)	.046	1.367	58%
8th Grade Read	55	(.112, -.005)	.117	n/a	71%
8th Grade Math	55	(.138, .068)	.070	2.027	69%
03-05 4th Read	18	(.107, .016)	.091	6.870	89%
05-07 4th Read	23	(.066, .065)	.001 (p=.49)	1.010	52%
03-05 8th Read	21	(.080, -.031)	.111	n/a	81%
05-07 8th Read	22	(.112, .021)	.091	5.314	52%
03-05 4th Math	18	(.173, .133)	.040 (p=.13)	1.302	63%
05-07 4th Math	22	(.128, .082)	.046 (p=.10)	1.559	64%
03-05 8th Math	21	(.118, .030)	.087	3.880	76%
05-07 8th Math	22	(.104, .079)	.025 (p=.14)	1.314	59%

Note: Unless otherwise noted, discrepancies were found to be significant at p<.01

Figure 4.5 Summary of state-NAEP trend discrepancies by subject and grade

For subject-grade-year combinations, the largest average state-NAEP discrepancy (0.111 units) was found for 8th grade reading in 2003-05, while the smallest average discrepancy (0.001) was found for 4th grade reading in 2005-07. Furthermore, while three of the four average discrepancies in reading were found to be significant, only one of the four math discrepancies (2003-05 8th grade math) showed a significant state-NAEP trend discrepancy. These statistically insignificant discrepancies may be due to the small sample sizes in these analyses.

Summary

The V statistics estimated for both state and NAEP trends indicate that state trends were significantly higher than NAEP trends in reading and mathematics at grades 4 and 8 from 2003-05, 2005-07, and 2003-07. In fact, average state V statistics were found to be more than twice as large as average NAEP V statistics. This provides motivation to explore potential explanations of these discrepancies.

The data show that state-NAEP trend discrepancies were higher for reading than for mathematics. This could possibly be explained by potentially larger content differences among state reading (or English Language Arts) tests and NAEP reading tests. Trend discrepancies were also found to be larger for 8th grade than for 4th grade, although no obvious explanation for this is available.

Once again, it is important to note that the significant differences between state and NAEP trends may be due to a combination of many factors, including differences in test content, examinees, examinee motivation, and educator manipulations. Before investigating the relationship between test security policy quality and state-NAEP discrepancies, the technical quality of the V estimates will be addressed.

V Estimates Compared to Other Measures of Trends

If it were possible, it would have been easier to estimate state-NAEP trend discrepancies by comparing traditional effect sizes, such as Cohen's d:

$$d = \frac{\bar{X}_2 - \bar{X}_1}{s_{\text{pooled}}} = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{s_1^2 + s_2^2}{2}}},$$

which, in this study, would be defined as the difference between mean scores at Time 1 and Time 2 divided by a pooled standard deviation (Cohen, 1988). Unfortunately, many states do not report this information, so effect sizes could not be calculated.

For the 15 cases in which traditional effect sizes (d) could be calculated, Figure 4.6 displays the relationship between these d values and the estimated V statistics. For these 15 cases, the average d effect size was found to be 0.0448 and the average V effect size was found to be 0.0618. Thus, the average traditional effect size was found to be smaller than the average V statistic by 0.0170.

For these 15 cases, the correlation between d and V estimates displayed in Figure 4.6 was found to be 0.729, indicating a strong positive linear relationship. The single obvious outlier, labeled on the figure, was for Delaware 8th grade reading from 2005-07. For this case, d was -0.028 and V was 0.209. Eliminating this outlier, the correlation increases to 0.971 and the difference between d and V estimates shrinks to 0.001. It is not known why Delaware is an outlier, although the means and standard deviations could have been misreported.

The V and d statistics can also be compared for NAEP trends. Figure 4.7 displays this relationship. Once again, a strong linear relationship exists, with a correlation of 0.977. The difference between the average d (0.0479) and the average V (0.0489) was only 0.001. This provides reassurance that the V estimates do, in fact, provide an effect-size measure of score trends.

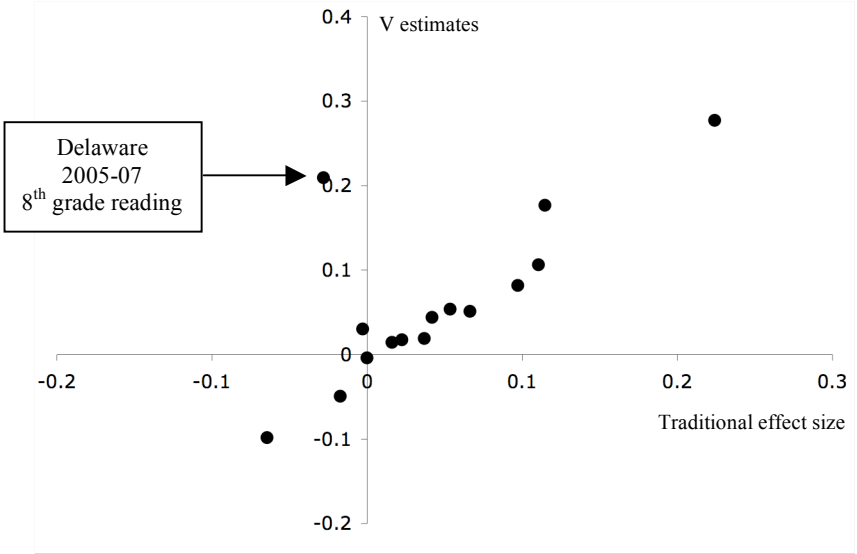


Figure 4.6 Relationship between *d* and *V* effect size estimates for state trends

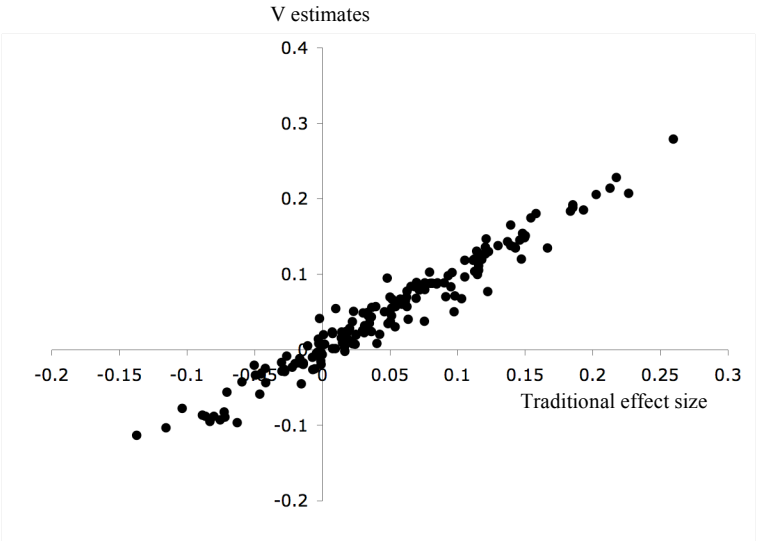


Figure 4.7 Relationship between *d* and *V* effect size estimates

The results from Figures 4.6 and 4.7 show that the V statistics used in this study are similar to traditional effect sizes. Now, the relationships between V statistics and other statistics are briefly investigated. First, Figure 4.8 displays the relationship between the V statistics and the change in the percentage of students scoring proficient on state tests.

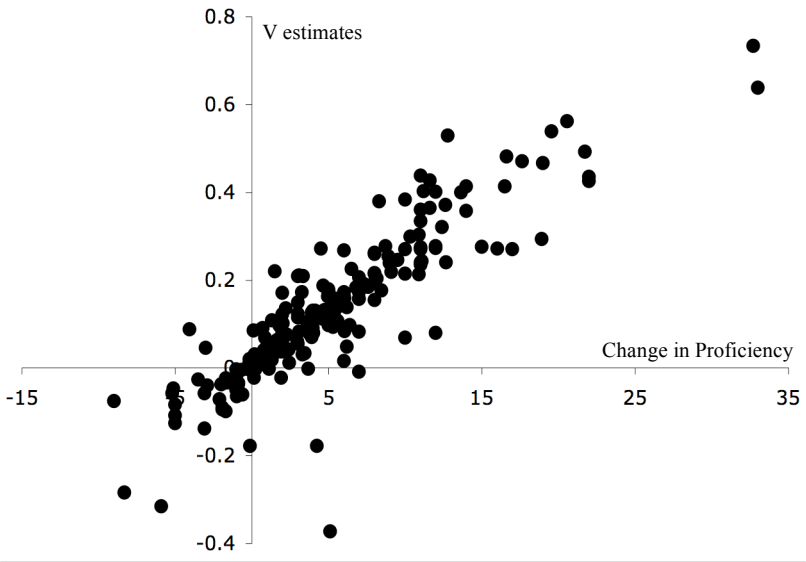


Figure 4.8 Relationship between proficiency trends and V discrepancies

As was discussed earlier, the change in percentage of proficient students depends on the choice of cut-score for proficiency. Thus, the scatterplot in Figure 4.8 should not be expected to display a perfect relationship. Nevertheless, a positive relationship is reassuring. The correlation between proficiency trends and V statistics was found to be 0.880.

In an attempt to provide further evidence of the shortcomings of the change in proficiency statistics, the observations falling in the second and fourth quadrants of Figure 4.8 represent cases with sign reversals in proficiency trends and V statistics. In

the figure, 2% of the observations represent cases in which proficiency declined yet *V* statistics indicated a positive score trend. 4% of the observations represent cases in which proficiency increased yet *V* statistics indicated a negative score trend. This could possibly provide evidence that educators are focusing instructional resources on students closest to proficiency (thus increasing proficiency in the face of negative score trend effect sizes), although other explanations are possible.

Choice of 3 Cut-Score Minimum

Recall that one potential limitation of the *V* statistic was with respect to its estimation procedure. For this study, a 3-cut-score minimum was chosen (states must report data corresponding to at least 3 cut-score to be included). Perhaps 3 cut-scores are too few for accurately interpolating P-P plots and estimating *V* statistics. One concern is that perhaps results would differ if a 4-cut-score minimum were chosen for this study.

To investigate the impact of choosing a 3-cut-score minimum, the *V* statistics estimated from states reporting 4 cut-scores were set aside. Then, for each of these states, the first cut-score was eliminated and *V* was re-estimated using the remaining 3 cut-scores. Likewise, *V* was estimated from the 3 cut-scores remaining after eliminating the second, third, and fourth reported cut-scores.

Figure 4.9 displays two examples of this analysis for 2 of the 184 observations. For each example (Kansas 2005-07 4^{th} grade mathematics on top; California 2003-05 4^{th} grade reading on bottom), the top row of graphs display interpolated P-P plots. The first P-P plot on the left was interpolated from all 4 reported cut-scores. The next four P-P plots were interpolated from 3 cut-scores after eliminating the first, second, third, and fourth cut-score data. The second row of graphs display the same information with the axes rotated and magnified to better show the interpolation results.

Below the graphs, the estimated area under the curve (Simpson's Rule) and *V* statistics are displayed in tables. Below these tables and to the right, a summary of the impact of reducing to 3 cut-scores is displayed.

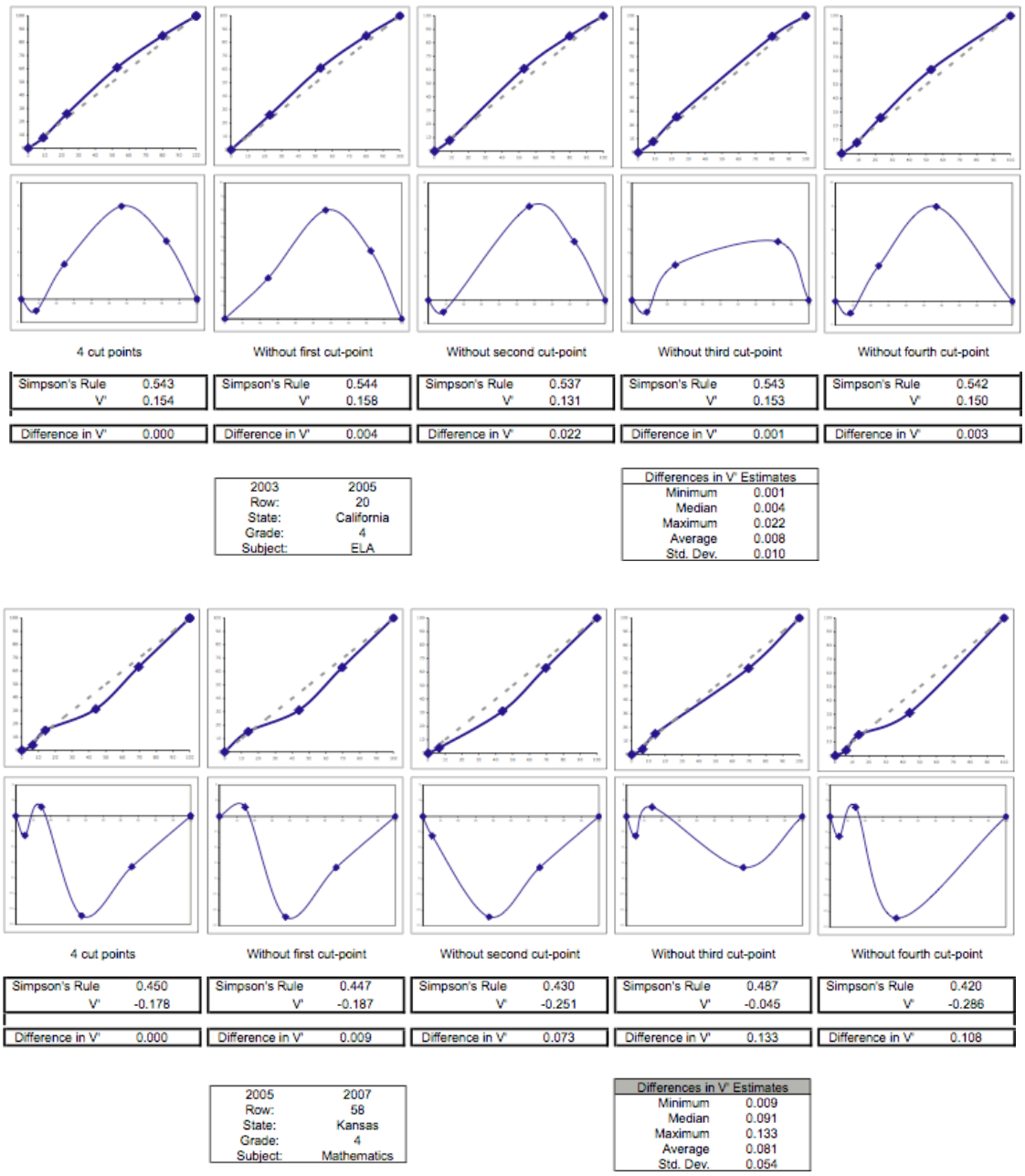
For the first (California) example in Figure 4.9, the *V* statistic was estimated to be 0.154 from all 4 cut-scores. After eliminating the first cut-score, the *V* statistic was estimated to be 0.158 (an absolute difference of 0.008). Eliminating the second, third, and fourth cut-scores led to absolute differences in estimated *V* statistics of 0.022, 0.001, and 0.003.

For the second (Kansas) example in Figure 4.9, the *V* statistic was estimated to be -0.178 from all 4 cut-scores. Eliminating the cut-scores one-at-a-time led to absolute differences of 0.009, 0.073, 0.133, and 0.108. The absolute difference of 0.133 was the largest difference found in this analysis.

For all 184 observations in this analysis, the absolute differences due to eliminating one cut-score ranged from 0.000 to 0.133, with an average absolute difference of 0.016. 60% of absolute deviations were less than 0.01 and 78% were less than 0.02. 3% of absolute deviations were found to be greater than 0.10.

While these differences appear to be small, recall that the average discrepancies between state and NAEP *V* estimates for 2003-05 and 2005-07 were 0.084 and 0.040. The average absolute difference of 0.016 (due to using 3 cut-scores instead of 4) represents 19% of the average discrepancy from 2003-05 and 40% of the discrepancy from 2005-07. Thus, even this apparently small absolute change in *V* estimates can have a significant impact on state-NAEP discrepancies.

The choice of which cut-score to eliminate also had an impact on the results. Removing the first cut-score had little impact on *V* estimates (an absolute difference of 0.006). Removing the second cut-score caused the *V* estimates to change by an absolute difference of 0.014. The third and fourth cut-score had the biggest impact. Their removal caused *V* estimates to change by 0.023 and 0.022, respectively.



	Minimum	25th %ile	Median	75th %ile	Maximum	Mean	Std. Dev
Absolute Deviation	0.000	0.003	0.009	0.017	0.133	0.016	0.024

Figure 4.9 P-P plots and V estimates after eliminating one cut-score

Instead of looking at absolute differences, the sign and magnitude of the changes in V estimates can be examined under the elimination of cut-scores. Removing the first cut-score caused V estimates to change by an average of -0.003. Removing the second and fourth cut-scores similarly caused V estimates to decrease by -0.001 and -0.013, respectively. Removing the third cut-score caused V estimates to increase by 0.008 units, on average. Overall, the average difference in V estimates was found to be -0.002. This would represent 2% and 5% of the state-NAEP discrepancies from 2003-05 and 2005-07.

In total, 81 of the 184 (44%) observations experienced an increase in V estimates after eliminating a cut-score, while 100 (56%) observations experienced a decrease in V estimates. While the magnitude of the difference in V estimates in moving from 4 to 3 cut-scores may be somewhat small, the fact that the estimates may either increase or decrease is a bit troubling. The results do show that the number of cut-scores reported by states does have an impact on the estimated trend effect sizes.

Summary: State-NAEP Trend Discrepancies

The analyses in this section indicate that the V statistics do provide effect-size estimates of score trends for both state tests and NAEP. Furthermore, the analyses found significant discrepancies between state and NAEP trends, with state trends approximately twice as large as NAEP trends. These results were replicated in both 4^{th} and 8^{th} grades; in both reading and mathematics; and for 2003-05, 2005-07, and 2003-07.

Figure 4.10 displays a map of the average state-NAEP discrepancy for each state (discrepancies were averaged for 2003-05 and 2005-07). The map shows that only five states (Alabama, Arizona, Massachusetts, West Virginia, Wisconsin) reported trends smaller than NAEP trends (with an average discrepancy of -0.04). All other states reported trends larger than NAEP trends (with an average discrepancy of 0.09). The map shows discrepancies are widespread and provides motivation to explore potential

explanations of these discrepancies. The next section investigates the relationship between test security policy quality and state-NAEP discrepancies.

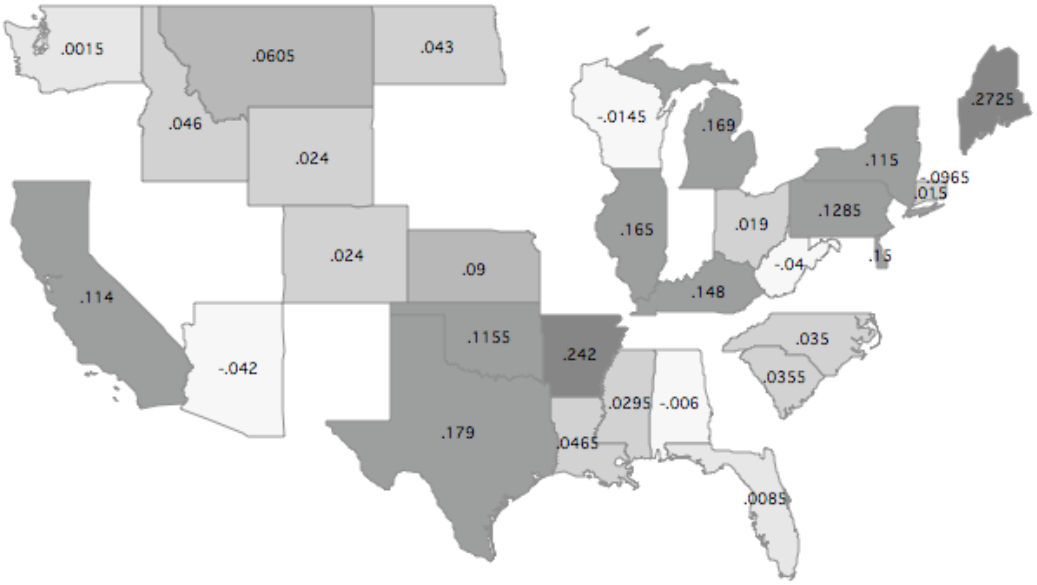


Figure 4.10 Average state-NAEP discrepancies for each state (2003-05, 2005-07)

Relationship Between Security Policy Quality and State-NAEP Score Trend Discrepancies

To investigate the relationship between the quality of state test security policies and the estimated discrepancies between state and NAEP trends, the discrepancy estimates for each state are regressed on the composite evaluation score for each state policy. Discrepancy estimates from 2003-05, 2005-07, and 2003-07 are analyzed separately as replications of the analysis. Figure 4.11 displays scatterplots of the relationships between policy quality and discrepancy estimates for each time period.

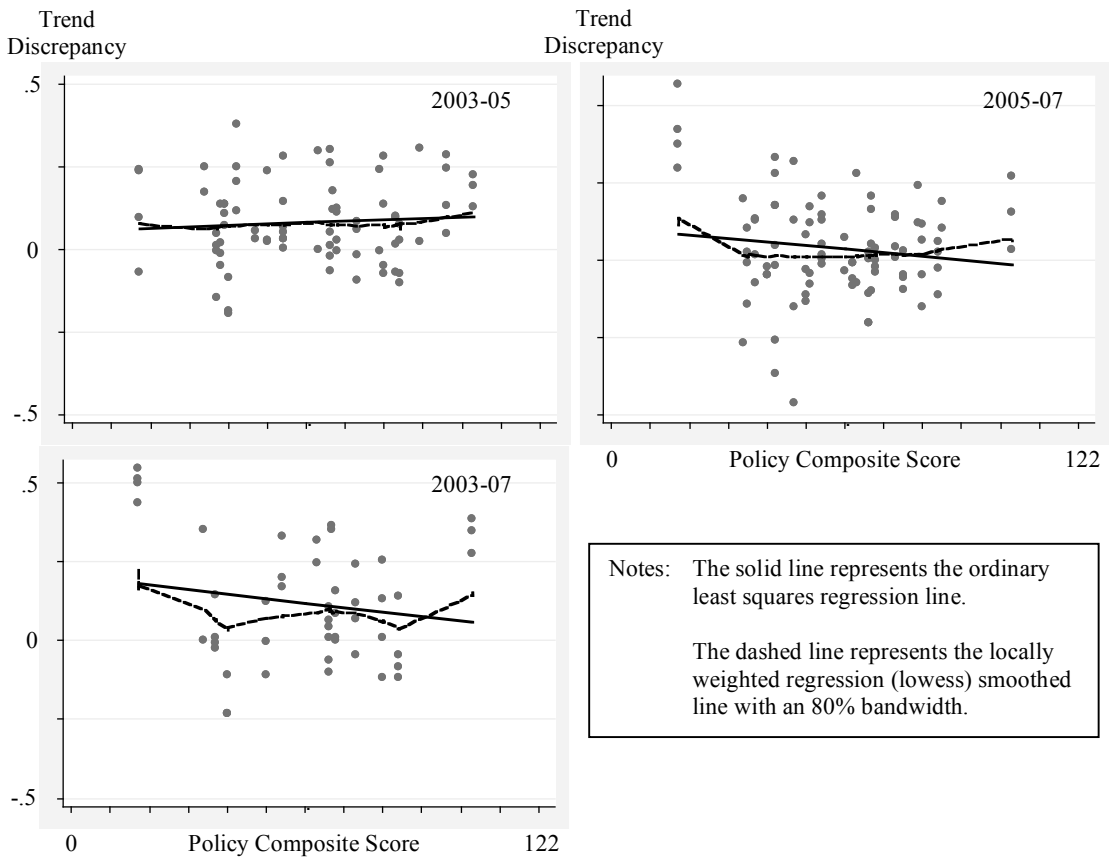


Figure 4.11 Scatterplots of policy quality composite scores and discrepancy estimates

Figure 4.11 shows no apparent relationship between policy quality and state-NAEP score discrepancies for any of the time periods. To aid in visualizing any relationships, two lines are drawn on each scatterplot in Figure 4.11. The solid line represents the ordinary least squares regression line and the dashed line represents the locally weighted regression (lowess) running-mean smoothed line.

If it were assumed that the relationship between the variables were linear, then the solid lines in Figure 4.11 show a slight positive relationship for 2003-05 and slight negative relationships between the variables in 2005-07 and 2003-07. The dashed lines, which do not assume a linear relationship, show that a linear relationship may hold for the 2003-05 data. The lines show the relationship for 2005-07 and 2003-07 are not linear, but perhaps quadratic relationships, with the worst and best policies experiencing

the greatest trend discrepancies. The main result to be taken from Figure 4.11 is that no apparent (or consistent) relationship between the variables is displayed.

To further investigate the relationship between policy quality and trend discrepancies, Tables 4.13 and 4.14 display correlations between the policy quality components and score trend discrepancies for each time period. Because the data yielded by the policy evaluation form cannot be assumed to have an interval scale and because the underlying distributions of the policy component scores cannot be assumed to follow a normal distribution, Spearman's rank-order correlations are calculated (Conover, 1999). The correlations show the strength of the relationship between the ranked quality of state test security policies and the ranked estimates of state-NAEP trend discrepancies. Correlations with magnitudes significantly greater than zero (at .05 and .01 levels of significance) have been highlighted in the tables.

The tables show that only 19 of the 300 (6.3%) correlations were found to significantly differ from zero, with 11 of those statistically significant correlations positive and 8 negative. In addition to the significance of each correlation, the consistency of correlations across the three time periods is important. If a relationship exists between a security policy component and trend discrepancy estimates, that relationship should be consistent across replications from 2003-05, 2005-07, and 2003-07. In Tables 4.13 and 4.14, boxes have been drawn around consistent correlations (that is, correlations with the same sign over all three replications) to highlight them. In the tables, 35 of the 100 sets of correlations were found to be consistent.

Table 4.13 Rank-order correlations

	Overall	Read	Math	G4	G8
COMPOSITE SCORE	.0568 -.0550 -.0969	.3940 -.0821 -.0401	-.1935 -.0694 -.1630	.0423 .0768 .0472	.0628 -.1736 -.2819
Formalize beliefs about the role of testing & testing practices	-.0123 -.0075 -.0667	.3262 -.0194 -.0288	-.2646 -.0245 -.1076	-.0191 .1516 .1456	-.0167 -.1521 -.3264
Prominence / Availability of information	-.0770 -.0853 **-.314**	.1676 -.0579 -.3090	-.2844 -.1737 -.3394	-.0827 .0925 -.1165	-.0990 -.2767 **-.574**[a]
Content	**.263** -.1647 .0548	**.427**[a] -.1686 .0662	.1711 -.1660 .0392	.2091 -.1399 .1013	.3013 -.1783 -.0585
Implementation	.1761 .0744 .1982	**.489**[a] .0578 .2664	-.0440 .0820 .1131	.1873 .1391 .2358	.1633 .0437 .1823
Requirements and sanctions	-.2075 .1055 -.1985	.0933 .0458 -.1237	**-.446**[a] .1380 -.2720	-.2089 .2820 .0062	-.2006 -.0677 **-.408**
Other	-.0507 -.0956 -.1136	.2538 -.0719 -.2198	-.2884 -.1362 -.0454	.0065 -.0817 -.0410	-.1170 -.0978 -.1600
Oversee test preparation, administration, and scoring activities	.0093 -.0249 -.0106	**.335** -.0374 .0938	-.2329 -.0536 -.1233	-.0197 .0892 .1181	.0336 -.1405 -.2112
Test Security Audits	.1284 .0963 .2164	.2368 .1807 .2294	.0825 .0023 .2002	.0713 .2376 .2919	.1680 -.0390 .0334
Test administration oversight	.1989 -.0036 .2824	**.328** -.0506 .3714	.1143 .0176 .1681	.1717 .0867 .3095	.2110 -.1012 .2012
Statistical Analyses	-.0707 -.0172 -.1920	.1889 .0225 -.1246	-.2682 -.1131 -.2607	-.0878 .0957 -.0291	-.0606 -.1310 **-.427**
Score Reports	.0623 .0513 .2084	.2434 -.0204 .2980	-.0950 .0963 .1258	.1029 **.331** **.447**	.0505 -.2440 -.0695

Notes: Values in each cell represent (from top to bottom) correlations for 2003-05, 2005-07, and 2003-07

Bold values represent correlations with $p < .05$

Values marked with [a] represent correlations with $p < .01$

So 65 of the 100 sets of correlations in the table were inconsistent. For example, the correlation between the composite policy evaluation score and estimated trend discrepancies was found to be .0568 from 2003-05. The correlations for the same two variables were found to be negative (-.0550 and -.0969) when data from 2005-07 and

2003-07 were used. While none of these correlations significantly differ from zero, it is still troubling how many correlations change signs over replications.

Table 4.14 Rank-order correlations

	Overall	Read	Math	G4	G8
Inform educators about why some activities are unacceptable	.1110 -.1880 -.0643	**.374** -.1954 -.0656	-.0946 -.2131 -.0751	.1309 -.1600 -.0039	.0806 -.1978 -.0888
Principles & Rules	.0183 -.1470 -.0003	.1230 -.2312 -.1498	-.0734 -.0851 .1337	.0827 -.2094 .0548	**.642**[a] -.0611 -.0424
Examples of appropriate and inappropriate behaviors	.2055 **-.221** -.2109	**.435**[a] -.1831 -.2197	.0131 -.2558 -.1952	.2224 -.1580 -.0548	.1898 -.2482 -.3238
General Guidance	-.1146 -.1649 -.2718	.0981 -.2090 -.2349	-.2764 -.1315 -.2830	-.1388 -.1280 -.2987	-.1145 -.1608 -.1993
Training	.1851 -.0089 .1181	**.464**[a] .0579 .2366	-.0243 -.1218 -.0214	.1567 -.0611 .0116	.1933 .0460 .2028
Limit opportunities for educators to manipulate test scores	.1653 -.0694 -.1514	.2791 -.1146 -.3197	.0657 -.0291 .0036	.1096 .0127 -.0207	.2075 -.1312 -.3500
Materials security	.1393 -.0538 -.0643	.2167 -.0379 -.2445	.0742 -.0790 .0921	.0108 .0361 -.1062	.2346 -.1596 -.1337
Test Forms	.0477 -.1198 **-.323**	.2336 -.2149 -.2809	-.1055 -.0462 -.3541	.0802 -.1037 -.1465	.0184 -.0998 **-.460**

Notes: Values in each cell represent (from top to bottom) correlations for 2003-05, 2005-07, and 2003-07

Bold values represent correlations with $p < .05$

Values marked with [a] represent correlations with $p < .01$

The combination of a small number of statistically significant correlations and inconsistent correlations across replications seem to indicate a lack of a relationship between test security policy quality and state-NAEP trend discrepancies. In fact, very few of the policy subcomponents have correlations that may *potentially* indicate some relationship with trend discrepancies. These subcomponents that are worthy of brief discussion are implementation, test security audits, general guidance, and score reports.

Looking at Table 4.13, the *implementation* subcomponent of test security quality (falling under the formalize component) shows a consistent pattern of positive correlations with trend discrepancies. 14 of the 15 replications across time periods, grade levels, and subjects show positive correlations. State policies scoring high in the implementation subcomponent were those that identified individuals responsible for test security, provided evidence that the policy was being implemented at the district and school levels, provided standard forms and checklists to aid school in implementation, and updated policy content regularly. The positive correlations would suggest that states with policies scoring higher in this subcomponent experienced larger state-NAEP trend discrepancies. While 14 of the 15 correlations between implementation and trend discrepancies were positive, it is important to note that only one of those correlations (for reading from 2003-05) was significantly higher than zero.

The *test security audit* subcomponent (located under the oversight component) also showed consistent correlations with trend discrepancies across replications, with 14 of the 15 replications showing positive correlations. This subcomponent simply rated the extent to which states audited the implementation of the statewide policy at the state-, district-, and school-levels. Even though the correlations were significantly positive, none of the correlations significantly differed from zero. Therefore, these correlations could be positive because of sampling error.

The *score reports* subcomponent (also located under the oversight component) rated the extent to which state policies described procedures educators could use to modify potentially inaccurate score reports. While only 11 of the 15 replications show consistently positive correlations, two of the three replications for fourth grade were found to be significantly greater than zero. Correlations between the score reports subcomponent and state-NAEP discrepancies were found to be significantly positive for the 2005-07 and 2003-07 trends. Unfortunately, these results did not generalize to other grade levels, across subjects, or over all three time periods.

The *general guidance* subcomponent (organized under the inform component) perhaps measures the most direct aspect of test security policies as they attempt to deter manipulations. Under this subcomponent, state policies were rated as to the extent to which they specifically informed educators that they cannot give students answers, modify student answer sheets, read specific sections of the test aloud to students, or provide students access to forbidden materials. While no statistically significant correlations between this general guidance subcomponent and trend discrepancies were found, 14 of the 15 correlations were consistently negative. This could indicate some potential relationship between this subcomponent and state-NAEP trend discrepancies.

Regression Analyses

While the Spearman's rank order correlations found no apparent relationships between policy quality and trend discrepancies, regression analyses provide another way to investigate potential relationships. If it is assumed that policy evaluation scores (composite, component, and subcomponent scores) follow normal distributions and that the scores are on an interval-scale, then ordinary least squares regression analyses can be used to look for any significant relationships.

The scatterplots (a sample of 3 are displayed in Figure 4.11) of policy evaluation scores and trend discrepancy estimates seem to show no clear functional relationships, so linear regression analyses were conducted. Table 4.15 displays the R^2 values, p-values, and root means square errors for a series of linear regression analyses. In each analysis, the dependent variable is the estimated state-NAEP trend discrepancy. The analyses were conducted separately for the 2003-05, 2005-07, and 2003-07 periods.

The first row of Table 4.15 (row A) displays the results when trend discrepancies were regressed on the composite policy evaluation scores. As the table shows, the composite scores accounted for 0.5%, 1.9%, and 2.5% of the variance in trend

discrepancies for 2003-05, 2005-07, and 2003-07. Thus, composite scores did not show a significant relationship with trend discrepancy estimates.

Row B of Table 4.15 displays the results when the four policy evaluation components were used as independent variables in the regression analysis. The components did not account for a significant proportion of the variance in discrepancy estimates in 2003-05 or 2005-07, but did account for almost 21% of the variance in discrepancies from 2003-07.

Table 4.15 Regression analyses and coefficients of determination

Independent variables in model:	R^2	p-value	Root MSE
A. Composite	.0057	.5251	.12667
	.0191	.3197	.16413
	.0253	.4190	.19579
B. Formalize, Oversee, Inform, Limit	.0400	.4820	.12700
	.1036	.0532	.15968
	.2077	.0412*	.18258
C. Formalize subcomponents	.2617	.0022*	.11215
	.1527	.0291*	.15168
	.1406	.1042	.19240
D. Oversee subcomponents	.0508	.2329	.12628
	.0088	.7422	.16791
	.1350	.0013*	.19077
E. Inform subcomponents	.2047	.0013*	.11559
	.0860	.2508	.16124
	.2740	<.0001*	.17478
F. Limit subcomponents	.0132	.6816	.12703
	.0638	.2518	.16128
	.1874	.0300*	.18075
G. All subcomponents	.3806	.0017*	.11069
	.1965	.1565	.16218
	.8072	<.0001*	.10280

Notes: Values in each cell represent (from top to bottom) results for 2003-05, 2005-07, and 2003-07

To further show how results did not replicate across time periods, Table 4.16 displays the standardized regression coefficients when trend discrepancies were regressed on the four component scores. The table shows that the standardized coefficients for the formalize, inform, and limit components changed signs from one time period to the next. Only the oversee coefficients remained positive for all replications. Also, the relative magnitudes of the standardized coefficients changed across replication. For 2003-05, the formalize coefficient had the largest magnitude. For 2005-07, the inform component yielded the largest coefficient. For the 2003-07 data, the limit component had the largest magnitude.

Table 4.16 Standardized regression coefficients

	2003-05	2005-07	2003-07
Formalize	-.3014	.1564	.0909
Oversee	.1459	.0767	.1539
Inform	.1985	-.2577	.2102
Limit	.1112	-.2070	-.6663

Other regression analyses showed similar inconsistencies across replications. While some significant relationships were found, these relationships held for only one or two of the time periods. Rows C-F of Table 4.15 display the results when all subcomponents under each of the four evaluation components are used to predict trend discrepancy estimates. The formalize subcomponents accounted for significant proportions of variance in trend discrepancies for 2003-05 and 2005-07, but not for 2003-07. The oversee and limit subcomponents each only showed a significant relationship with discrepancies from 2003-07. The inform subcomponents showed significance for only the 2003-05 and 2003-07 trend discrepancies.

Row G of Table 4.15 displays the results when all subcomponents under all four components were entered into the regression analysis. Once again, while the independent variables did account for a significant proportion of the variance in trend discrepancies for 2003-05 and 2003-07, they did not account for a significant proportion from 2005-07. The subcomponents did account for more than 80% of the variance in trend discrepancies from 2003-07, but less than 20% of the variance from 2005-07.

As a final attempt to find significant relationships between test security policy quality and trend discrepancies, stepwise regression analyses were conducted. For each time period, the trend discrepancies were regressed on all 73 items from the policy quality evaluation form. Unfortunately, the results were again inconsistent across time periods. While the existence of a frequently asked questions section on a state's test security web page did account for 31% of the variance in trend discrepancies from 2003-05, this significant result did not hold for the other time periods. Likewise, the clarity of a policy seemed to predict score trend discrepancies in 2005-07 and the use of independent test administration monitors seemed to best predict trend discrepancies from 2003-07.

Before moving onto to further analyses, one more regression was conducted. State-NAEP trend discrepancies were regressed on the number of published news report on test score manipulations. For the 2003-05 discrepancies, the numbers of published news reports from 2000-2003 were used as the independent variable. For the 2005-07 trend discrepancies, the independent variable was the number of news reports from 2003-2005. For the 2003-07 discrepancies, the number of published reports from 2003-07 served as the independent variable. For all three analyses, the number of published news reports accounted for less than 1% of the variance in trend discrepancy estimates.

Reasons For Lack of Significant Relationships

The correlation and regression results reported earlier all support the conclusion that state test security policy quality does not significantly and consistently predict state-NAEP score trend discrepancies. Both technical and substantive reasons could help explain why no consistent relationships were found. First, it may very well be the case that no relationship exists between these variables. Second, if a relationship did exist, it may only appear in school- or district-level data. The fact that individual educators (teachers, principals, or even superintendents) may choose to manipulate test scores does not mean these manipulations would necessarily have a big impact on statewide test results.

Some technical reasons why significant and consistent relationships were not found include the lack of a linear relationship between these variables and the impact of range restriction on the results. If it were the case that security policies and trend discrepancies had a nonlinear relationship, then these linear regression analyses would not necessarily discover this relationship. Unfortunately, no clear nonlinear relationship between the variables could be found from visual inspection of scatterplots. Also, recall that the states included in these analyses tended to be the states with the higher policy scores (see Table 4.1). The fact that states with lower policy scores were not included could restrict the range of the independent variable and make finding a significant relationship more difficult.

Categorical Analyses

Recall that state test security policies were placed into seven different sets of categories: clear vs. ambiguous; state-level vs. district level; punitive vs. instructive; independent monitoring vs. no independent monitoring; investigative vs. preventative; example-based vs. not example based; and positive message vs. negative message. It was

hypothesized that if test security policies had a relationship with trend discrepancies, then differences should exist between the groups in each of these seven categorizations.

Before testing to see if group differences exist, group sample sizes, means, and standard deviations are examined. Table 4.17 displays these summary statistics for states within each categorization.

Table 4.17 Summary of trend discrepancy estimates by categorizations

	2003-05	2005-07	2003-07
Clear & Accessible	n=57; $\bar{X}$=.090; s=.124	n=69; $\bar{X}$=.029; s=.155	n=35; $\bar{X}$=.103; s=.154
Ambiguous/Hard-to-find	n=21; $\bar{X}$=.066; s=.134	n=20; $\bar{X}$=.078; s=.195	n=13; $\bar{X}$=.164; s=.284
Unclassified	n=0	n=0	n=0
State-level	n=52; $\bar{X}$=.075; s=.131	n=46; $\bar{X}$=.012; s=.156	n=24; $\bar{X}$=.071; s=.183
District-level	n=11; $\bar{X}$=.142; s=.125	n=24; $\bar{X}$=.107; s=.203	n=9; $\bar{X}$=.261; s=.257
Equal state/district	n=11; $\bar{X}$=.077; s=.095	n=15; $\bar{X}$=.041; s=.097	n=11; $\bar{X}$=.149; s=.145
Unclassified	n=4; $\bar{X}$=.057; s=.145	n=4; $\bar{X}$=-.040; s=.108	n=4; $\bar{X}$=.012; s=.054
Punitive (laws/sanctions)	n=30; $\bar{X}$=.065; s=.125	n=36; $\bar{X}$=.040; s=.094	n=20; $\bar{X}$=.079; s=.171
Instructive/Informative	n=38; $\bar{X}$=.093; s=.129	n=43; $\bar{X}$=-.010; s=.159	n=22; $\bar{X}$=.082; s=.162
Unclassified	n=10; $\bar{X}$=.104; s=.125	n=10; $\bar{X}$=.257; s=.220	n=6; $\bar{X}$=.390; s=.202
Independent monitoring	n=23; $\bar{X}$=.122; s=.116	n=25; $\bar{X}$=.051; s=.125	n=17; $\bar{X}$=.172; s=.156
No independent monitoring	n=55; $\bar{X}$=.068; s=.128	n=64; $\bar{X}$=.036; s=.178	n=31; $\bar{X}$=.091; s=.212
Investigative	n=31; $\bar{X}$=.045; s=.131	n=35; $\bar{X}$=.043; s=.102	n=21; $\bar{X}$=.048; s=.151
Preventative	n=37; $\bar{X}$=.113; s=.115	n=50; $\bar{X}$=.007; s=.168	n=23; $\bar{X}$=.119; s=.175
Both equally	n=2; $\bar{X}$=.165; s=.201	n=0	n=0
Unclassified	n=8; $\bar{X}$=.076; s=.123	n=4; $\bar{X}$=.417; s=.114	n=4; $\bar{X}$=.497; s=.046
Example based	n=30; $\bar{X}$=.091; s=.143	n=35; $\bar{X}$=.028; s=.140	n=20; $\bar{X}$=.070; s=.199
Not many examples	n=44; $\bar{X}$=.083; s=.120	n=50; $\bar{X}$=.044; s=.185	n=28; $\bar{X}$=.155; s=.190
Unclassified	n=4; $\bar{X}$=.030; s=.020	n=4; $\bar{X}$=.091; s=.097	n=0
Positive message	n=28; $\bar{X}$=.075; s=.110	n=35; $\bar{X}$=-.003; s=.156	n=18; $\bar{X}$=.073; s=.166
Negative message	n=36; $\bar{X}$=.071; s=.131	n=46; $\bar{X}$=.057; s=.116	n=26; $\bar{X}$=.093; s=.168
Both positive & negative	n=2; $\bar{X}$=.165; s=.201	n=0	n=0
Unclassified	n=12; $\bar{X}$=.131; s=.139	n=8; $\bar{X}$=.135; s=.336	n=4; $\bar{X}$=.497; s=.046

From these summary statistics, inconsistent results can be eliminated. For example, the first row of Table 4.17 displays the summary statistics for the states placed into *clear and accessible*; *ambiguous or difficult-to-find*; or *unclassified* categories. The table shows that 57 state-NAEP trend discrepancy estimates from 2003-05 came from states with *clear and accessible* policies. Likewise, 21 state-NAEP trend discrepancies from 2003-05 were estimated from states with *ambiguous or difficult-to-find* policies. The states in the *unclassified* category did not provide data to estimate state-NAEP trend discrepancies, so the sample size is zero in this analysis.

In this row, the table shows that from 2003-05, states with *clear and accessible* policies had an average state-NAEP trend discrepancy estimate of 0.90. States with *ambiguous or difficult-to-find* policies experienced an average trend discrepancy estimate of only 0.66. Thus from 2003-05, *clear and accessible* policies were related to larger trend discrepancies.

Before running any formal test of the statistical significance of this group difference, the consistency of the findings are examined by looking at the summary statistics for 2005-07 and 2003-07. In this example, the table shows that the *clear and accessible* policies were related to larger discrepancies from 2003-05 and 2005-07; but smaller discrepancies from 2005-07. Since the direction of the group differences changed across replications, no formal hypothesis testing procedures will be conducted. This categorization of state policies showed no consistent relationship with state-NAEP trend discrepancies across the three time period replications.

Table 4.18 displays which categorizations yielded consistent or inconsistent results. The table shows that results from only three different categorizations showed consistent results across the 2003-05, 2005-07, and 2003-07 replications. These categorizations were the *state-level vs. district-level*, the *independent monitoring vs. no independent monitoring*, and the *positive vs. negative message* policy categorizations.

Now that consistent results have been found for three different categorizations, more formal hypothesis testing can be investigated. For each categorization with consistent results, the hypothesis to be tested would state that the groups do not differ in trend discrepancy estimates. These hypotheses could be tested for each row using a simple analysis of variance if the normality, equal variances, and independence assumptions are met.

Table 4.18 Consistency of group differences across time periods

Categorization	Consistency of results across 2003-05, 2005-07, and 2003-07 periods
Clear & Accessible Ambiguous/Hard-to-find Unclassified	Inconsistent group differences across replications
State-level District-level Equal state/district	District-level policies consistently show higher trend discrepancies than state-level policies.
Punitive (laws/sanctions) Instructive/Informative	Inconsistent group differences across replications
Independent monitoring No independent monitoring	Policies requiring independent monitoring of test administration show consistently higher trend discrepancies.
Investigative Preventative	Inconsistent group differences across replications
Example based Not many examples	Inconsistent group differences across replications
Positive message Negative message Unclassified	States with unclassified policies consistently show higher trend discrepancies than states with policies with clear positive or negative messages.

Based on visual inspection of the distributions of trend discrepancies for each time period, it seems reasonable to assume that state-NAEP trend discrepancies follow

normal distributions. Further supporting this conclusion, Shapiro-Wilk tests for normality found that the normality assumption cannot be rejected at a 0.10 level of significance for any of the time periods (Shapiro & Wilk, 1965).

Based on an inspection of the standard deviations in Table 4.17 and variance ratio tests, the equal variance assumption does not appear to be reasonable for some of these tests. Due to the heterogeneity of variances across groups and unequal sample sizes across groups, the parametric analyses of variances should not be conducted (Conover, 1999). Unfortunately, since the nonparametric Kruskal-Wallis test assumes identically-shaped distributions for each group (Conover, 1999), the differences in variances also eliminates its use.

The independence assumption is also unreasonable to make for these tests. Since each state can have four trend discrepancies in each time period (grades 4 and 8 in reading and mathematics) and it can be assumed that state-level factors (curriculum, student populations, test security policies) affect these trend discrepancies estimates, the observations cannot be assumed to be independent.

Because of the lack of independence among trend discrepancies within states, a groups-within-treatments design must be used. In this design, two levels of analyses are conducted. First, the minor units (trend discrepancy estimates for all states) are analyzed to determine if dependencies exist within each state. If significant dependencies are found, then the analysis must be conducted at the state level. If no significant dependencies within states are found, then the analysis of variance can be conducted with the minor units to determine if significant differences exist between groups.

Tables 4.19 and 4.20 display the results of these analyses to determine if significant differences in trend discrepancies exist between states with state-level policies and those with district-level policies. Table 4.19 displays the number of and average trend discrepancies for each state falling in each policy category in each time period. The table shows that while the average discrepancies for state-level policies are smaller than

the average discrepancies for district-level policies for each time period, the states falling within each category vary wildly.

To determine if the analysis should be conducted on the individual trend discrepancy observations (minor units) or on the average state discrepancies (major units), Table 4.20 displays summary tables for the groups-within-treatments analysis. The tables show the amount of variation in trend discrepancies due to the groups (state- vs. district-level policies; row 1); variation in trend discrepancies due to the states within each group (between states within treatments; row 2); variation within each state (within states; row 3); and unexplained error variance for the minor units analysis (error; row 4).

Table 4.19 Average trend discrepancies for state- and district-level policies

	2003-2005		2005-2007		2003-2007	
Category	State	Mean (Obs.)	State	Mean (Obs.)	State	Mean (Obs.)
District	Colorado	0.096 (3)	Alabama	-0.006 (4)	Colorado	0.004 (3)
	Illinois	0.165 (2)	Arkansas	0.242 (4)	Kansas	0.176 (2)
	Kansas	0.214 (2)	Colorado	-0.048 (4)	Maine	0.497 (4)
	Maine	0.128 (4)	Kansas	-0.034 (2)	**Average**	**0.261 (9)**
	Average	**0.142 (11)**	Maine	0.417 (4)		
			North Dakota	0.043 (4)		
			Ohio	0.019 (2)		
			Average	**0.107 (24)**		
State	California	0.167 (3)	Arizona	-0.042 (4)	California	0.233 (3)
	Connecticut	-0.023 (4)	California	0.061 (3)	Connecticut	0.029 (4)
	Delaware	0.119 (2)	Connecticut	0.053 (4)	Massachusetts	-0.190 (3)
	Hawaii	0.044 (2)	Delaware	0.181 (2)	Oklahoma	0.356 (2)
	Idaho	0.239 (4)	Hawaii	-0.041 (4)	Pennsylvania	0.281 (2)
	Massachusetts	-0.155 (3)	Idaho	-0.147 (4)	South Carolina	0.070 (4)
	Montana	0.030 (4)	Kentucky	0.148 (2)	Washington	0.004 (2)
	New York	0.115 (4)	Massachusetts	-0.038 (3)	Wisconsin	-0.027 (4)
	North Carolina	0.035 (4)	Montana	0.091 (4)	**Average**	**0.071 (24)**
	Oklahoma	0.149 (2)	Oklahoma	0.082 (4)		
	Pennsylvania	0.151 (2)	Pennsylvania	0.106 (2)		
	South Carolina	0.074 (4)	South Carolina	-0.003 (4)		
	Texas	0.179 (4)	Washington	-0.154 (2)		
	Washington	0.157 (2)	Wisconsin	-0.013 (4)		
	Wisconsin	-0.016 (4)	**Average**	**0.012 (46)**		
	Wyoming	0.024 (4)				
	Average	**0.075 (52)**				

Table 4.20 Groups within treatments analysis for state vs. district level policies

Source (2003-05)	Sums of Squares	Degrees of Freedom	Mean Squares	Mean Square Ratio
State/district level policies	0.041	1	0.041	1.708 [b] (p=.196)
Between states within treatments	0.498	18	0.028	3.928 [a] (p<.001)
Within states	0.303	43	0.0070	
Error (between + within states)	1.485	61	0.024	

Source (2005-07)	Sums of Squares	Degrees of Freedom	Mean Squares	Mean Square Ratio
State/district level policies	0.142	1	0.142	4.733 [b] (p=.033)
Between states within treatments	1.493	19	0.079	7.088 [a] (p<.001)
Within states	0.543	49	0.0111	
Error (between + within states)	2.036	68	0.030	

Source (2003-07)	Sums of Squares	Degrees of Freedom	Mean Squares	Mean Square Ratio
State/district level policies	0.236	1	0.236	5.021 [b] (p=.032)
Between states within treatments	1.021	9	0.114	5.897 [a] (p<.001)
Within states	0.423	22	0.0192	
Error (between + within states)	1.445	31	0.047	

Notes: a) The mean square ratio represents the ratio of between states within treatments and within states

b) Mean squares ratios assuming the trend discrepancies within states are independent

If the trend discrepancy estimates within each state were independent, the tables would show that the variation between states within groups would be relatively small. To test the relative magnitude of this source of variation, the ratio of *between states within groups* and *within states* mean squares is calculated. For all three time periods, this mean square ratio was found to be statistically significant at a 0.001 level. Thus, a significant *states within groups* effect was found and the trend discrepancy estimates within each state should not be assumed to be independent. The analyses will have to be conducted using the average state-level trend discrepancy estimates (major units).

Rather than conducting this *groups within treatments* analysis for the other two categorizations with consistent results, it is assumed that dependencies exist within state trend discrepancy estimates. So state-level (major unit level) analyses are conducted.

These analyses use the unweighted mean trend discrepancies for each state (ignoring the number of trend discrepancy estimates for each state). These unweighted means are displayed in Table 4.21.

Table 4.21 Unweighted mean state-level trend discrepancies for each categorization

	Unweighted Means		
Category	2003-05	2005-07	2003-07
State-level	0.1509	0.0904	0.2256
District-level	0.0808	0.0202	0.0943
Independent Monitors	0.1140	0.0806	0.1888
No independent monitors	0.0717	0.0366	0.0892
Positive message	0.0889	-0.0085	0.0840
Negative message	0.0771	0.0632	0.1170
Unclassified	0.1305	0.1348	(1 observation)

Table 4.22 displays the results of the analysis based on these unweighted means for each time period. As the table shows, no statistically significant differences were found between the state-level and district-level policy groups for any of the time periods. Thus, it cannot be concluded that states with state-level policies experience smaller trend discrepancies than states with policies implemented at the district-level.

Table 4.23 displays the results from a similar analysis of the differences between states requiring independent monitoring of test administration and states with no such requirement. Once again, no statistically significant differences were found between the groups. Thus, it cannot be concluded that states with policies requiring independent monitoring of test administration experience different trend discrepancies than states that do not require independent monitoring of test administration.

Table 4.24 displays the results from a similar analysis of the differences among states with policies having an overall positive message, states with policies having an overall negative message, and states with policies that could not be classified as positive

or negative. Once again, no statistically significant differences were found between the groups. Thus, it cannot be concluded that a relationship exists between the tone or message of a policy and state-NAEP trend discrepancies

Table 4.22 Groups within treatments analysis for state vs. district level policies

Source (2003-05)	Sums of Squares	Degrees of Freedom	Mean Squares	Mean Square Ratio
State/district level policies	0.0160	1	0.0160	1.883
Between states within treatments	0.1527	18	0.0085	p = .1869

Source (2005-07)	Sums of Squares	Degrees of Freedom	Mean Squares	Mean Square Ratio
State/district level policies	0.0230	1	0.0230	1.394
Between states within treatments	0.3135	19	0.0165	p = .2523

Source (2003-07)	Sums of Squares	Degrees of Freedom	Mean Squares	Mean Square Ratio
State/district level policies	0.0376	1	0.0376	0.9518
Between states within treatments	0.3557	9	0.0395	p = .3548

Table 4.23 Groups within treatments analysis for independent monitoring policies

Source (2003-05)	Sums of Squares	Degrees of Freedom	Mean Squares	Mean Square Ratio
Independent monitoring	0.0095	1	0.0095	1.1216
Between states within treatments	0.1952	23	0.0085	p = .3005

Source (2005-07)	Sums of Squares	Degrees of Freedom	Mean Squares	Mean Square Ratio
Independent monitoring	0.0107	1	0.0107	0.6903
Between states within treatments	0.3725	24	0.0155	p = .4143

Source (2003-07)	Sums of Squares	Degrees of Freedom	Mean Squares	Mean Square Ratio
Independent monitoring	0.0357	1	0.0357	1.1053
Between states within treatments	0.4193	13	0.0323	p = .3031

Table 4.24 Groups within treatments analysis for policy tone

Source (2003-05)	Sums of Squares	Degrees of Freedom	Mean Squares	Mean Square Ratio
Policy tone (positive/negative)	0.0067	2	0.0033	0.1003
Between states within treatments	0.6583	20	0.0329	p = .9050

Source (2005-07)	Sums of Squares	Degrees of Freedom	Mean Squares	Mean Square Ratio
Policy tone (positive/negative)	0.0490	2	0.0245	1.775
Between states within treatments	0.3718	23	0.0138	p = .1919

Source (2003-07)	Sums of Squares	Degrees of Freedom	Mean Squares	Mean Square Ratio
Policy tone (positive/negative)	0.0037	1	0.0037	0.1445
Between states within treatments	0.3066	12	0.0256	p = .7015

So as the previous results show, no consistent or significant differences were found in trend discrepancies among the seven different categorizations of state test security policies. This could mean that no relationships exist between state test security policies and state-NAEP trend discrepancies, or it could mean that the data collected in these analyses were not able to demonstrate the relationship. Some possible reasons for this include the small sample size within each categorization (due to the dependencies of trend discrepancies within states) or the lack of a representative sample of state policies within each group (because many states were excluded from the analyses due to a lack of trend discrepancy estimates).

Longitudinal Analyses

As the final analyses of this data, changes in state-NAEP trend discrepancies are compared over time for states that made significant changes to their test security policies. As was mentioned earlier, only five states can be included in this analysis: Hawaii, Louisiana, Mississippi, Pennsylvania, and Washington. Because of the small,

nonrepresentative nature of the sample, no hypothesis testing procedures are conducted. Instead, simple descriptive statistics are calculated.

Table 4.25 displays the average state-NAEP trend discrepancies for these five states in the periods before and after significant changes were made to their security policies. The table shows that four of the five states experienced smaller state-NAEP trend discrepancies after making modifications to their test security policies. Louisiana was the only state that experienced a higher trend discrepancy following its modification. While these results are intriguing, a simple comparison to states that did not make modifications to their policies shows that, overall, the average state reported smaller trend discrepancies in 2005-07 than they did in 2003-05. Thus, no generalizations are made from these descriptive statistics.

Table 4.25 Longitudinal Analyses

State	2003-05	Modification	2005-07	Difference
Washington	0.157	Added security budget in 2005	-0.154	-0.311
Hawaii	0.044	Heightened security in 2007	-0.041	-0.085
Mississippi	0.065	Statistical detection in 2006	-0.006	-0.071
Pennsylvania	0.151	Published Do's Don'ts guide in 2006	0.106	-0.045
Louisiana	0.008	Statistical detection in 2005	0.085	+0.077
All other states	0.089	No significant modifications	0.056	-0.033

Summary: Relationship Between State Test Security Policy Quality and State-NAEP Trend Discrepancies

The correlation and regression analyses in this section were unable to uncover any apparent, consistent relationships between the quality of a state's test security policy and the estimated discrepancies in score trends between the state test and NAEP. While state policies that explicitly state that educators cannot manipulate answer sheets or give students answers experienced slightly lower trend discrepancies, these relationships were

not statistically significant. Likewise, while the test security policy quality subcomponents were able to account for 38%, 20%, and 81% of the variance in trend discrepancies (for 2003-05, 2005-07, and 2003-07, respectively), the standardized regression coefficients changed in relative magnitude and sign from one time period to the next. These results all lead to the same conclusion – the quality of test security policies, as measured by the FOIL framework, has no consistent or significant relationship with state-level trend discrepancies.

The categorical analyses similarly failed to find significant differences in trend discrepancies among different groups of security policies. While states with state-level policies experienced consistently lower trend discrepancies than states with district-level policies; states with independent monitoring of test administration experienced lower trend discrepancies than states without independent monitoring; and states with clearly positive or negative messages experienced lower trend discrepancies than states with unclassifiable policies; these differences were not found to be statistically significant.

The lack of statistical significance could be due to the lack of a real relationship between test security policies and score trend discrepancies. It could also be due, in part, to a relatively small, nonrepresentative sample size in these analyses. These problems were due to a lack of available data to estimate trend discrepancies for some states, and due to the fact that trend discrepancy estimates within each state cannot be assumed to be independent.

Perhaps the most promising sign that a relationship between test security policy quality and score trend discrepancies exists comes from the longitudinal analysis. Although only five states could be included in this analysis, four of the five states that made significant changes to their test security policies experienced smaller trend discrepancies after making these modifications. The nature and size of this sample does not allow the results to generalize to all other states, but these results are nonetheless interesting.

CHAPTER 5: SUMMARY, DISCUSSION, AND RECOMMENDATIONS

Summary

Published news reports, along with results from surveys, direct observations, statistical analyses, and targeted research all indicate that some educators do engage in activities designed to increase student test scores without an equal, corresponding increase in student achievement on the underlying construct. These manipulations, were appear to be widespread and growing in use, destroy the validity of inferences made from test scores. Little research has investigated the methods educators use to manipulate test scores, the impact of these manipulations on test results, and the effectiveness of measures taken to deter these manipulations.

This study attempted to determine if a relationship exists between the quality of state test security policies and discrepancies between state test and NAEP score trends over three time periods from 2003-2007. To do this, a taxonomy of manipulations was developed to classify the methods of manipulations educators have used to increase test scores. This taxonomy was then used to gain a better understanding of the reasons why educators manipulate test scores and how the impact of these manipulations may be estimated.

To estimate the impact of these manipulations, discrepancies in score trends from state tests and NAEP results were estimated using a nonparametric framework. The framework yielded scale-invariant effect sizes for the trends in state test and NAEP scores. This framework was used because of the limitations of comparing percentages of students scoring proficient on each test over time. This framework was also used because many states do not report data that can be used to calculate traditional effect sizes.

This study then investigated the measures taken in an attempt to deter educators from manipulating test scores. After discussing the limitations of many of these

measures, state test security policies were investigated. A FOIL framework was developed to evaluate the quality of each state's test security policy with respect to how well the policy formalizes the beliefs of educators regarding testing, provides for oversight of testing activities, informs educators about appropriate and inappropriate behaviors, and limits the opportunities for educators to manipulate test scores. This framework was used to obtain test security policy quality scores for each state policy.

Finally, these test security policy quality scores were compared with the estimated state-NAEP trend discrepancies estimates to determine if a relationship exists. Analyses were conducted to determine which aspects of a state's test security policy seem to have the strongest relationship with score trend discrepancies. These relationships were investigated with the understanding that several plausible rival hypotheses could possibly explain why the relationships might exist.

Discussion

This study synthesized results from surveys, news reports, and research to discover that test score manipulations fall into one of four categories: manipulations of the teaching philosophy or process; manipulations of the examinee pool; manipulations of the test administration; or manipulations of score reports or score standards. The results indicate that up to 88% of educators have engaged in one of 65 different test score manipulations.

Further synthesizing results from other research, this study found four main reasons why educators manipulate test scores. First, educators are former students and research has shown that cheating on tests is widespread among high school and postsecondary students. If some educators experienced benefits from cheating as students, then there is reason to believe that these educators would continue to cheat as adults. Second, educators face increasing pressure from accountability systems. If some educators do not agree with the accountability movement, they may try to undermine it

(or game the system) by manipulating test scores. Third, many educators are unaware of which behaviors are appropriate or inappropriate. With the overabundance of professional codes and conflicting expert opinions about which behaviors are inappropriate, educators may feel it is up to them to determine which activities are appropriate. Fourth, some educators may manipulate test scores because of a lack of oversight and test security policies.

To prevent educators from manipulating test scores, some have recommended making changes to the tests used for accountability. Performance assessments (or tests with new items each year) would reduce the opportunity for educators to manipulate test scores, but the cost would be prohibitive. Others have recommended using statistical analyses to detect potential manipulations and punish educators found to have manipulated test scores. These methods would also be inadequate, given the inability of these statistical detection methods to accurately detect manipulations and ineffectiveness of punishments. To prevent educators from manipulating test scores, this study recommends that states develop high quality test security policies to provide oversight and guidance for all testing activities.

Based on expert recommendations and an analysis of the currently existing test security policies, this study recommends state test security policies have four main components. First, the policies should formally state the beliefs of educators with regards to testing. The policies should be clearly written, identify individuals responsible for test security, and outline sanctions faced by those found to have manipulated test scores. Second, the policies should provide for oversight of test preparation, administration, and scoring activities. The policy should be audited regularly, provide for independent monitoring of test administration, and provide for statistical analyses of answer sheets to detect potential manipulations. Third, the policies should inform educators about which behaviors are appropriate or inappropriate. To do this, the policies should provide specific examples of appropriate and inappropriate activities and should provide for

regular training of all personnel involved in testing. Fourth, the policies should limit the opportunities educators have to manipulate test scores. To do this, the policies must provide clear guidelines to ensure the security of test materials.

These recommendations were used in evaluating the test security policies currently implemented in each state. The evaluation found that the quality of policies varies greatly among states, with state policies earning between 3% - 84% of the possible number of points. Within the framework, states scored relatively highest in informing educators about manipulations. The states had the greatest opportunity to improve test security by focusing on the oversight of testing practices. Some specific recommendations to improve security, based on this evaluation, include:

- Provide more oversight into testing activities by:
  - Performing more test security audits at the school- or district-levels
  - Implement statistical analyses to detect potential manipulations
  - Provide for independent monitoring of test administration
- Provide a more formal statement of educators beliefs about testing:
  - Allow teachers to provide input into the content of the security policy
  - Explain the importance of test security
  - Outline the process used to investigate potential manipulations
  - Provide a standard form for educators to report potential manipulations
  - Provide protection for individuals who report potential manipulations
- Limit opportunities for educators to manipulate test scores by:
  - Limiting the amount of time educators have access to testing materials
  - Providing for multiple test forms or new test items each year
- Better inform educators about which behaviors are inappropriate:
  - Explain the importance of validity and how manipulations destroy validity
  - Provide clear guidelines for choosing test preparation activities
  - Provide regular training in test security and make materials available online

- Clearly state if educators can sanitize answer sheets before scoring

To determine if a relationship exists between state test security policy quality and test score manipulations, an estimate of the impact of test score manipulations must be developed. In this study, the discrepancies between score trends from state tests and NAEP tests were used as possible estimates of the impact of manipulations. Because manipulations increase test scores without increasing student achievement, and because NAEP tests are more resistant to manipulation, then discrepancies between state and NAEP results could provide an estimate of the extent to which state test results are inflated.

Comparisons of state and NAEP results for single-year or trends are limited if they are based on the percentages of students scoring above or below a specific cut-score. This is because these statistics depend on the choice of cut-score. For this study, a scale-invariant framework was used to provide effect sizes invariant with respect to the choice of cut-scores. These scale-invariant effect sizes showed that state test trends were significantly higher than NAEP trends in both reading and mathematics in both 4^{th} and 8^{th} grade from 2003-05, 2005-07, and 2003-07.

While these significant discrepancies between state and NAEP trends could provide an estimate of the impact of manipulations on state test scores, it is important to remember that several plausible rival hypotheses could also account for these discrepancies. State test and NAEP results might be expected to differ due to the impact of content differences, item format differences, test difficulty, score standards, test administration procedures, examinee motivation, or other factors.

Even though no causal relationship could be ascertained, the relationship between state test security policy quality and state-NAEP trend discrepancies was of interest. To investigate this relationship, correlational analyses were conducted on the policy evaluation and trend discrepancy data. These analyses were not able to find any significant relationships that held for the 2003-05, 2005-07, and 2003-07 time periods.

Regression analyses and comparisons among seven different categorizations of policies were also unable to find any apparent relationships.

Limitations

Technically, this study may have been most limited by the size and nature of the available data. While test security policies were evaluated for every state, these evaluations were based on policy document content. The quality of a state's test security does not only depend on the quality of its content, but also by how well it is implemented at the state-, district-, school-, and classroom-levels. In this study, the quality of policy implementation was based on documents auditing policy implementation at the district level. Because an accurate evaluation of policy implementation could not be completed, the state test security evaluations might not accurately reflect the actual quality of the test security policy procedures.

Another limitation dealt with the lack of documentation of modifications made to state test security policies. Many states did not clearly indicate when security policies were implemented or changed. In an attempt to document these modifications, policy documents were compared from one year to the next in order to look for changes. The magnitude and frequency of these changes were difficult to assess. In this study, it was assumed that policy quality remained constant over the entire 2003-2007 time period unless major changes to policy content were discovered. This assumption may be tenuous.

Yet another limitation of the policy evaluation surfaces because of the subjective nature of the evaluation framework. While the evaluation rubric was clarified after examining all state policy documents (to better define what was meant by a score of 0, 1, or 2 on each item), the rubric had no a priori definitions for these scores. Thus, an evaluation of state test security policies completed at a different time (or by a different individual) might produce slightly different scores for each state.

The size and representativeness of the sample was also limited by the lack of available data to estimate state-NAEP trend discrepancies. Many states reported test results with fewer than 3 cut-scores. Other states changed tests or cut-scores, or simply did not report data that could be used to estimate these scale-invariant effect sizes. If this missing data could be assumed to be missing at random, the impact may not be very large. Unfortunately, the states from which trend discrepancies could not be estimated tended to be the states with the lower policy evaluation scores. The extent to which this excluded data impacted the results of the study is unknown.

The trend discrepancy estimates could also have been impacted by the decision to require data to be reported from a minimum of 3 cut-scores. Perhaps requiring a larger number of cut-scores would have improved the accuracy of the interpolation function, but it would have been at the expense of excluding even more states from the study.

This study was also limited in analyzing data from the 2003-05, 2005-07, and 2003-07 time periods in grades 4 and 8 in reading and mathematics. These grades, subjects, and time periods were selected because of their correspondence with NAEP tests. Thus, any generalizations of the results from this study to other grade levels, subject areas, or tests should be made with caution.

This study may have also been limited by its focus on state-level discrepancies and policies. It may be the case that the impact of test score manipulations may only be detected at the classroom-, school-, or district-levels. It may also be the case that test security policies, even though they are developed and implemented at the state level, might only show an impact for individual districts, schools, or classrooms. Thus, the lack of a relationship between test security policy quality and trend discrepancies may be due to the fact that such a relationship may only be detectable at a lower level.

It might also be the case that no relationship exists between the quality of state test security policies and state-NAEP score trend discrepancies. The final set of limitations in this study deal with the plausible rival hypotheses that may explain why

state test score trends are significantly more positive than NAEP trends. As stated earlier, these discrepancies may be due to a variety of test content, item format, test administration, and/or examinee factors. In order to determine which factors influence these discrepancies, the other plausible factors must be eliminated.

Recommendations

In addition to the policy recommendations made earlier for state test security and the recommendation that states report test score means and standard deviations (to calculate effect sizes), recommendations for further research are presented. The methods, frameworks, and results from this study could be used to inform future research into test score manipulations, test security policies, and state-NAEP trend discrepancies.

Since this study did find significant discrepancies in trends between state and NAEP tests, attempts to discover the reasons for these discrepancies are recommended. Several researchers have studied the effects of content differences (Wei, Shen, Lukoff, Ho, & Haertel, 2006), examinee motivation (Klein, Hamilton, McCaffrey, & Stecher, 2000; Linn, Baker, & Betebenner, 2002), examinee demographics, test item formats, and test administration differences (Jacob, 2007), but no research has systematically attempted to eliminate plausible rival hypotheses for these differences. More research into potential causes of state-NAEP trend discrepancies is recommended.

More research into test security policies is also recommended. While research has concluded that policies to prevent student cheating are effective (McCabe & Trevino, 2002), very little research has been conducted into the effectiveness of policies designed to prevent test score manipulations (Cizek, 1999, p. 171). This study found that the quality of policy content varies widely from state-to-state over the four components of test security. More research into the implementation of these policies, especially at the district- and school-levels, would further provide further evidence of the effectiveness of test security policies. Case studies of policy implementation at the state-, district-, and

school-levels over time would provide a great deal of information that could be used to improve test security policies. The policy evaluation framework developed in this study is also recommended as a framework for states to audit their test security policies and procedures

In addition to the impact on test score manipulations, the impact of test security policies on educator morale is of interest. Future research could investigate public and educators' perceptions of test security policies. Research could also be conducted to determine the relative cost of developing and implementing test security policies at the state-level.

Finally, more research into the impact of test score manipulations is recommended, especially at the district- or school-levels. While comparisons of score trends between state and NAEP tests provide useful information at the state level, perhaps more appropriate audit tests could be employed at the district- or school-levels. The results from these audit test comparisons, teacher surveys, and aberrant response analyses could be combined to gain a better understanding of the impact of test score manipulations. Combining this with information about policy implementation at the school-level might provide a better measure of the relationship between test security policy quality and test score manipulations.

APPENDIX A: PUBLISHED NEWS SUMMARIES

- 12/21/2007 USA Today:

  Doris Alvarez, principal of Preuss high school in San Diego, resigns in connection with a case of alleged cheating and grade-tampering. An audit by the University of California at San Diego found that three-fourths of reviewed Preuss transcripts had one or more grades changed — most of them to benefit students. Just last month, the school, which prepares low-income and minority students for college, ranked 10th out of 18,000 U.S. high schools in new rankings by U.S. News & World Report. Preuss ranked second in the USA among charter high schools. Last May, Preuss ranked 10th among 1,200 high schools deemed the nation's best by Newsweek.
  (Toppo, 2007)

- 12/16/2007 Orlando Sentinel:

  In June, a Pasco County teacher was fired for rewording questions while proctoring the math portion of the FCAT. The Caveon Test Security company, which investigates cheating allegations and has a contract with Florida, estimates that adult-led cheating is probable in some 1 percent or 2 percent of the schools it has investigated.
  (Julian, 2007)

- 12/13/2007 New York Post:

  A former Staten Island school administrator ordered teachers to change scores on state Regents exams to enable students to graduate - and even hiked some of the test scores herself at home. Former Wagner HS Assistant Principal Mary Incantalupo, of Staten Island, was recommended for termination for her role in a 2006 test-tampering scandal that's been under investigation by the Department of Education since October 2006. The school's principal, Gary Giordano - who was dating Incantalupo at the time of the grading and who married her this year - was cleared of serious wrongdoing, but will be disciplined for not keeping tabs on the exams.
  (Gonen, 2007)

- 11/11/2007 KSTP Eyewitness News (Minnesota):

  A Channel 5 Eyewitness News investigation finds teachers help their students cheat on standardized tests. According to reports from the Minnesota Department of Education, teachers may find it tempting to point out a correct answer on a multiple choice test or correct a composition prior to grading. According to the Minnesota Board of Teaching, one teacher had their license suspended for three years for altering student's answers on a test. Another teacher had their license suspended for nine months for administering a test improperly. Even though the test booklets are supposed to be sealed and kept locked up, the reports showed that teachers were found to be taking the tests, sharing them with other teachers, and in one case, a teacher's pre-test lesson plan included a math problem which was "strikingly similar" to one appearing on the actual test.
  (Muehlhausen, 2007)

- 11/01/2007 Boston.com:

  20 Massachusetts teachers are accused of improperly helping students during the MCAS exams in 2007. This compares to 15 accusations in 2006 and 3 in 2005. A New Bedford elementary school teacher briefed students on the subjects of the reading passages they would encounter on the reading MCAS. Other teachers provided dictionaries or other forbidden tools, or made mistakes in administering the exam, such as forgetting to remove helpful material from a visible place in the classroom.
  (Jan, 2007)

- 10/15/2007 The Columbus Dispatch:
  An analysis of investigations in Ohio's 10 largest school districts in 2006 finds that 5 of the 8 educators accused of cheating are either under investigation by the state or have been punished already. Some districts prefer to let the problem teachers move on quietly to another school, a practice known as "passing the trash." Some districts have a "basics-only" policy for reference checks that prevents them from telling potential employers anything except when the teacher worked for the district.
  (Smith Richards & Riepenhoff, 2007)

- 10/12/2007 The Detroit News:
  Thousands of fifth and sixth grade students in Michigan will be forced to retake the writing portion of the Michigan Educational Assessment Program test after a newspaper published sensitive information about the test. The Jackson Citizen Patriot published two of the writing topics before the test administration period. The decision to retest students was made after the Michigan Department of Education learned a reporter was allowed into Jackson Public Schools during the administration of the test, a violation of the state's testing ethics.
  (Mrozowski, 2007)

- 09/07/2007 The Press-Enterprise (California):
  An English teacher at Citrus Hills Intermediate School in Corona, CA and a fourth-grade teacher at Red Maple Elementary School in Moreno Valley, CA are caught allegedly giving their students copies of old exams as practice for the current year's test. The two schools where these teachers worked were among 15 California schools that are currently under investigation for testing irregularities. This is not the first report of cheating in these school districts. A math teacher at Sierra Middle School in Riverside Unified District resigned in 2004 after being caught changing student answers on dozens of state tests.
  (Parsavand, 2007)

- 08/23/2007 Recordnet.com (California):
  A school district in Calaveras reports allegations of teacher cheating to state officials. One teacher in the district allegedly read portions of the state test or taught content during the test administration. Another teacher allegedly used actual test questions to prepare students for the test.
  (Johnson, 2007)

- 08/23/2007 CBS 13 (California):
  One teacher at August Elementary school and another at McKinley Elementary school in Stockton, California are accused of cheating to increase their students' scores on the state tests. The McKinley school teacher was caught writing down questions from the test, while the August teacher is accused of giving a student "punctuation tips." District official Dianne Barth is quoted as saying, "I don't believe it's widespread. I don't believe there is cheating in Stockton Unified [school district]."
  (Cannata, 2007)

- 08/23/2007 Washington Post:
  Severna Park High School in Washington D.C. is put on probation after allegations of cheating on an Advanced Placement history exam. A College Board investigation found that the test proctor failed to follow test directions and allowed students to talk and use cell phones. 42 students were forced to re-take the test and the test proctor was banned from ever administering the AP tests again.
  (Wan, 2007)

- 08/18/2007 The Modesto Bee:

Twelve schools in San Joaquin Valley, California submit 13 reports of testing irregularities to state education officials. One teacher at Oak View School allegedly made a practice worksheet of items "almost identical" to the items on the CAT-6 exam. The school's principal suggested the behavior was due to the pressure felt by teachers under No Child Left Behind to increase test scores. Another teacher allegedly allowed her students to use calculators on the math test. A third alleged incident involves a teacher who helped students during the writing exam. Another teacher was caught copying answers from student test booklets.
(Balassone, 2007)

- 08/16/2007 Dayton Daily News:
  Test scores from the City Day Community charter school in Dayton, Ohio plummet when independent officials monitor the 2007 test administration. In 2006, the school had produced extraordinary score gains under suspicion that students were given practice tests that were identical to the actual test.
  (Elliott, 2007b)

- 08/10/2007 Herald Tribune (Florida):
  Mary Cropsey, a third-grade teacher at Mills Elementary School in Manatee, Florida, is accused of tampering with student answer sheets on the Florida Comprehensive Assessment Test (FCAT). One student reports that Cropsey helped students on the test; another student reported hearing that the teacher gave students extra time to complete the exam. An investigation began after yet another student reported that she had not finished the exam, but the next day all the bubbles had been filled-in. If the allegations are proven true, Cropsey could lose her teaching certificate and even be charged with a crime.
  (Morris, 2007)

- 07/16/2007 Miami Herald:
  Hollywood Hills Elementary, an A-rated school for the past five years, receives an incomplete grade because FCAT scores for 16 students were flagged as irregular. The students in Katie Steinberg-Lessard's third grade class do not have their test scores months after taking the test. The teacher spent three mornings a week before school with students who wanted extra practice for the FCAT. Five other Dade schools with incomplete grades are currently being investigated.
  (Shah, 2007)

- 07/16/2007 San Francisco Chronicle:
  The California Department of Education concludes that for the second consecutive year, educators at University Preparatory Charter High School in San Francisco interfered with state-mandated testing. State investigators seized illegal copies of the 2005 form of the test that was used to prepare students for the exams. Eight former teachers at the school assert the existence of a culture of cheating at the school. According to those former teachers, student grades are frequently falsified and low-scoring students are excluded from state-mandated testing. Last year, the state found that hundreds of answers on the ninth-grade English and math tests had been changed from wrong to right. A counselor from Oakland's Skyline High school reports that a student earning D's and F's transferred to University Preparatory Charter High School and received A's and B's while taking 16 classes in a single semester. When the student returned to Skyline High, he once again earned D's and F's. Last year, investigators concluded that educators at the school changed hundreds of test answers before they were sent for scoring. Former testing coordinator Mike Schwartz is suing school founder and director Isaac Haqq for breach of contract, claiming Haqq was responsible for the altered answer sheets.

  Eight days after the original story was published, Isaac Haqq resigned as principal from University Preparatory Charter High School.
  (Asimov, 2007ab)

- 07/13/2007 The Dallas Morning News:
  A state investigation finds that David Tamez, an elementary school teacher in Amarillo, Texas, leaked the fourth-grade writing test prompt on the spring TAKS writing test to colleagues before the test administration. Tamez reportedly leaked the test information because he believed educators in other districts were doing it as well. The teacher obtained the test information by volunteering to serve on the committee that selects questions for the final form of the TAKS. He alleges that committee members "regularly smuggle out secret TAKS information to share in the home districts." Another teacher interviewed by investigators signed a statement indicating that Tamez "bragged that the source of his insider test information was… a person he had sex with who works for a company that helps build the TAKS." The Amarillo Independent School District concluded that the teacher obtained the information from an unidentified employee at Pearson Educational Measurement. Tamez resigned from his position, but will retain his teaching certificate if he cooperates with the investigation.
  (Benton, 2007b)

- 07/03/2007 News 10 Now:
  An investigation concludes that a third grade teacher at West Leyden Elementary in the Adirondack Central School District in Boonville, NY cheated for students on the New York State Mathematics Exam. According to parents, the teacher told students how she was going to help them and then tapped on any incorrect answers during the test administration. According to superintendent Frederick Morgan, the teacher remains employed by the district – she was "… moved to a grade where there is no state exam given."
  (Ohler, 2007)

- 06/27/2007 New York Times:
  A 23-month investigation into cheating allegations at Cobble Hill High School of American Studies in Brooklyn, NY concludes that the whistle-blower wrongly accused both the principal and assistant principal. Teacher Philip Noble accused principal Lennel George and assistant principal Theresa Capra of ordering teachers to cheat on the scoring of Regents exams. The investigation alleges that Mr. Noble was "a sub-par teacher with poor evaluations who wrongly accused [them] of engineering a cheating scheme because [they] had given him a negative review that could have led to his firing." The investigation does not explicitly rule out the possibility of cheating at the school; only that the principal and assistant principal did nothing wrong.
  (Bosman, 2007)

- 06/24/2007 New York Times:
  A 2006 investigation concludes that wrong answers were erased and changed to correct answers on state-mandated English tests in four New York City elementary schools. 2007 test scores for the schools in which cheating allegedly took place fell substantially, providing evidence that cheating had inflated the 2006 scores.
  (Fessenden, 2007)

- 06/24/2007 Newsday.com:
  State officials blame "adult interference" for suspiciously high test scores in eight schools in Camden, NJ. Faculty members reportedly allowed students to use calculators, which were not allowed on the exam.
  (Marcus, 2007a)

  The entire Uniondale school district is placed on academic probation due to evidence of tampering with Regents Math A and B high school exams and the State Mathematics Assessments for grades 3-8 in 2005 and 2006. The New York Department of Education reports that complaints of test fraud have more than doubled over the past five years, with the department receiving 37 complaints in 2006. One dozen teachers and administrators accused of test fraud have faced hearings in front of the New York Professional Standards and Practices Board. Of those twelve

cases, six cases resulted in revocation of professional certifications, two cases were cleared, and the remaining four cases remain under investigation. The number of complaints verified by the state has remained relatively steady, with between 9-16 in each of the past five years.
(Hildebrand, 2007a)

An analysis of Uniondale's test scores found that 333 answers on the Regents Math A exam were altered, and 97% of the time they were changed to the correct answer. On the Regents Math B exam, 198 answers were changed, with 97% again being changed to the correct answer. On the 2005 8th grade math assessment, Uniondale students scored below average on 11 of the 14 easiest questions, but higher than average on 12 of the 13 most difficult items.
(Hildebrand, 2007b; Marcus, 2007b)

- 06/21/2007 KLTV-7:
  Lincoln Intermediate schoolteacher Bernice Martin allegedly changed answers on 17 math TAKS tests in 2006. The investigation into the alleged test fraud included erasure and handwriting analyses. The teacher, a counselor who served as test coordinator, and a retired principal could lose their credentials over the incident.
  (McCollum, 2007)

- 06/15/2007 The Dallas Morning News:
  The Texas Education Agency could begin proceedings to close Theresa B. Lee Academy, a Fort Worth charter school, do to alleged tampering with the 2005 administration of the Texas Assessment of Knowledge and Skills (TAKS). The school was identified through a statistical analysis of test scores conducted by test security firm Caveon. When the state followed-up on this analysis, the school repeatedly refused to provide information to investigators. When asked for testing paperwork, principal William Powell reportedly replied that the paperwork had been, "lost in the flood." The details of this flood story changed with each telling and state investigators were unable to confirm the existence of any flood. The academy's vice principal, Shirley Dukes, reportedly changed answers on student answer sheets, informed teachers of the test's essay topics before test administration, and even wrote essays for the students. Dwaine Guyton, a former teacher at the academy, alleges that vice principal Dukes changed student answer sheets after test administration. Jakobus Wolf, a former science teacher at the academy, alleges that vice principal Dukes copied test answers onto the chalkboard for students and later asked if Mr. Wolf would be interested in being paid to manipulate student answer sheets. Mr. Wolf is reported to have said, "Kids know that if they go to Theresa B. Lee, somebody else will pass the TAKS for them."
  (Benton, 2007a)

- 06/14/2007 Texas Education Agency News:
  The Texas Education Agency is recommending sanctions against three educators in three schools because of cheating on the TAKS. An analysis found excessive erasures and evidence of tampering with student answer sheets at Winona High School. In San Augustine Intermediate School, a student complained that someone had changed his answers on the 7th grade math test. An analysis later found evidence of tampering on the answer sheets of 17 out of 25 students in that class.
  (Texas Education Agency, 2007a)

- 06/12/2007 WREG-TV Memphis:
  Connie Smith is fired over a cheating scandal in Tunica County, Mississippi. The teacher at Robinsonville Elementary witnessed the school's principal leaking test answers and reported the misconduct to the district superintendent. The school board terminated Smith's contract even though two other teachers confirmed her account of the misconduct.
  (Turner, 2007)

- 06/06/2007 Newsday:

New York state auditors discover 11 schools opened exam materials prematurely. Under state rules, the exam materials must be stored in steel safes or concrete vaults and are not to be unsealed until the day of testing. At 14 other New York schools, auditors found that exams had been removed from their locked boxes before being stored in safes. The state Education Department has recently revoked the rights of more than 20 schools to store exams after finding security breaches.
(Hildebrand, 2007c)

- 06/04/2007 The Dallas Morning News:
A conservative statistical analysis of 2005-2006 TAKS answer sheets conducted by Dr. George Wesolowsky, a professor at McMaster University in Canada, finds that the scores from more than 50,000 students show evidence of cheating that could include students copying answers from other students or educators doctoring student answer sheets. The analysis found 112 schools in which at least 10% of answer sheets were flagged for cheating. Many of the suspicious scores were found on the 11th grade test – the test students must pass to graduate. The schools with the strongest evidence of cheating include Forest Brook High School (North Forest ISD); Worthing High School and Sam Houston High School (Houston ISD); and South Oak Cliff High School (Dallas ISD). Based on the analysis, it appears as though cheating was more than 3 times as common in Dallas and Houston as it was in other large Texas districts. Professor Wesolowsky is quoted as saying, "The evidence of substantial cheating is beyond any reasonable doubt."
(Benton & Hacker, 2007a, 2007b)

- 05/26/2007 Visalia Times-Delta:
Three Visalia Unified School District teachers in Visalia, CA are reported for misconduct during the 2006 administration of tests for the Academic Performance Index (API). One La Joya Middle School teacher and one Crestwood Elementary School teacher are reported to have read portions of the test that were designed to be read by students. A teacher at Mineral King Elementary reviewed questions with students after administering the test. District Superintendent Stan Carrizosa reportedly views the incidents as innocent mistakes.
(Garcia, 2007)

- 05/22/2007 WMC-TV Memphis:
A teacher at Germanshire Elementary in Memphis, Tennessee allegedly cheated for students on the TCAP test. The district's evidence of this misconduct includes erasure marks on test booklets, stray marks on answer sheets, and statements from students.
(Rhodes, 2007)

- 05/20/2007 Inside Bay Area:
Three high school teachers in Oakland resign after being caught cheating on the California High School Exit Exam. Two of the teachers reportedly clarified a test question on the math portion of the exam while the third teacher proctored the exam. Dale Brodsky, an attorney hired by the Oakland Education Association, is quoted to have said cheating is "a non-issue in this whole debate about testing," and then questioned the term cheating by saying, "What is 'cheating'?"
(Murphy, 2007)

- 05/13/2007 San Francisco Chronicle:
Teachers in at least 123 public schools have reportedly cheated for students on California's high-stakes tests between 2004-2006. In two-thirds of these cases, the schools admit that they had cheated. The cheating behaviors included (a) allowing students to use reference materials such as maps and flow charts during the test, (b) allowing students to use calculators, (c) helping students answer questions, and (d) erasing and changing student answers. California currently identifies potential misconduct by scanning answer sheets for suspicious erasures. Cheating is virtually ignored in schools in which cheating impacts less than 5% of tests are given. Schools in which cheating impacts more than 5% of the tests are not ranked and receive a note stating "adult

irregularity in testing procedure" occurred. Since 2005, the following San Francisco area schools have confirmed testing irregularities: Mission Elementary (Antioch Unified); Hidden Valley Elementary, Cambridge Elementary, and Glenbrook Middle (Mount Diablo Unified); Forty-Niners Academy and Ravenswood City Elementary (East Palo Alto); John Muir Elementary (San Francisco Unified), Scott Lane Elementary (Santa Clara Unified), Bay Farm Elementary (Alameda City Unified), Chavez Middle and Treeview Elementary (Hayward Unified), Los Paseos Elementary (Morgan Hill Unified), Petaluma Junior High, Petaluma Joint Union High, Fair Oaks Elementary, and Williams Elementary (San Jose Unified).
(Asimov & Wallack, 2007)

- 05/11/2007 Associated Press:
  A teacher in Bloomington, Ohio is placed on paid leave after allegedly helping students cheat on the state graduation test. Nine students will be forced to retake the exam instead of graduating with their classmates.
  (Associated Press, 2007a)

- 05/02/2007 St. Petersburg Times:
  Barbara Heggaton, a special education teacher at Moon Lake Elementary School in Pasco County Florida, is accused of giving answers to three students during administration of the FCAT.
  (Solochek, 2007)

- 05/01/2007 Tyler Morning Telegraph:
  An investigation by the Texas Education Agency Office of Inspector General reports that test administration misconduct occurred in the 2005 administration of the TAKS in Winona High School. While the district was cleared of any wrongdoing in three previous investigations, the report alleges misconduct on the basis of suspicious erasure patterns and a possible breach of security caused by a missing key to the cabinet in which tests were kept.
  (Waters, 2007)

- 02/04/2007 Dayton Daily News:
  A newspaper investigation found that students at City Day Elementary School in Dayton, Ohio were given 44 practice questions that were identical or "substantially the same" as questions from the actual state exam. In some questions on the practice test, only names or small details were changed from the real test questions. The investigation was launched due to the suspiciously large amount of improvement shown by the school. In 2005, no sixth grade student in the school passed the math subtest of the Ohio Achievement Test. One year later, 100% of these students (now in 7^{th} grade) passed the math test.
  (Elliott, 2007)

- 12/03/2006 Deseret News:
  The Utah State Board of Education accepted a test protocol pamphlet that defines cheating on the U-PASS exams. According to state testing director Judy Park, the ethics policy comes as a response to, "an unusually high volume of calls to the state office from testing directors, parents, teachers, and superintendents with ethical questions on the way tests are given." According to Park, the state receives about five reported testing protocol violations each year. The new policy defines the following behaviors as cheating: (a) changing student answers in any way, for any reason, (b) looking at a test beforehand and altering lessons, (c) using inflections or gestures to help students in any way, (d) leaving helpful materials on classroom walls, (e) reclassifying student ethnicity, (f) letting students or parents supervise other students taking a test, (g) suggesting a student rethink his or her answer, (h) letting students take testing materials away from the testing site.
  (Toomer-Cook, 2007)

- 11/20/2006 New York Daily News:

City officials are investigating teachers from Millenium Art Academy in Castle Hill for allegedly coaching 35 students during testing and inflating student scores.
(Einhorn & Melago, 2006)

- 11/06/2006 Staten Island Advance:
Seventeen Staten Island teachers inform the United Federation of Teachers of tampering with the Regents exam. The vice principal at Wagner High School allegedly re-scored student tests at home while teachers added points to student test scores. The teachers claim they were told to change test answers in their classrooms. The informants also claim the principal said he would make them pay for coming forward. Other Staten Island teachers suggest this behavior is a system-wide practice. According to Frank DeSantis, a teacher in St. George High School, "A lot of teachers get that feeling that all [schools] are looking for is statistics, and [they're] lying and cheating to get them."
(Gonen, 2006; W-CBS TV, 2006)

- 10/22/2006 The Columbus Dispatch:
Of the 28 Ohio school districts analyzed by The Columbus Dispatch, 15 had instances of educators cheating on standardized tests. Barbara Oaks, a teacher in the Coventry district, looked through the test and wrote out a geometry problem she thought her students would have trouble with. Winifred Shima, a teacher from the Parma district, used a copy of the test to create a study guide for students that included 45 of the 46 actual test questions. Brian Wirick (East Knox) and Heather Buchanan (Wapakoneta) both used the test to create study guides for students. Judy Wray, a veteran teacher in Marietta, made copies of the actual state test to help students prepare. Wray is reported to have said that teachers cheat more than administrators know.
(Richards, 2006a)

- 10/11/2006 The Indianapolis Star:
Two Corpus Christi Catholic School teachers in South Bend, Indiana are found to have cheated on statewide exams. Beth Troyer and Sandra Ernst were suspended for one week without pay for allegedly sending questions and answers (from an older version of the test) home with the students. State officials have received about a dozen reports of testing violations this year, but only half are suspected cheating incidents.
(Hupp, 2006)

- 10/01/2006 The Dallas Morning News:
5 months after being found guilty for cheating on the Texas Assessment of Knowledge and Skills (TAKS), at least 10 of the 22 Wilmer-Hutchins teachers are now working in other North Texas Public Schools. More than two years after the cheating took place, none of the teachers ever faced official sanction. Several of the school districts that now employ these teachers were unaware that these teachers have cheated in the past.
(Benton, 2006b)

- 09/25/2006 The Indy Channel.com:
A fifth-grade teacher from Wayne Township, Indiana receives a one-week suspension without pay for allegedly giving four students extra time to complete the math portion of the Indiana State Test of Educational Progress. Tom Langdoc, the district's Director of School Community Services, believes the teacher was aware that she was cheating.
(The Indy Channel, 2006)

- 08/31/2006 Fairtest.org:
A teacher's aide and a guidance counselor at Morton Elementary in Franklin City, VA are suspended after allegedly changing student answers on state exams. School officials report the answer changes would have actually resulted in more student failures on the exam.
(Fairtest, 2006)

- 08/20/2006 The Boston Globe:
  The Massachusetts Department of Education documents 15 cases of inappropriate educator behaviors on the 2006 administration of the MCAS (compared to 3 allegations in 2005). A sixth-grade teacher from Andover West Middle School is reprimanded for reviewing a student's test and returning it to the student for revision. A fifth-grade test booklet at Pentucket Lake Elementary School was stolen and mailed to a local newspaper. Teachers in New Bedford and Peabody allowed students to use dictionaries during the test.
  (Jan, 2006)

- 07/30/2006 Houston Chronicle:
  Two Houston fifth-grade teachers resign after being accused of giving test answers to their students. Sheryle Douglas and Shawn Manning, the teachers once praised by President Bush and Oprah Winfrey, admit to giving students answers to an old version of the Stanford 10 Achievement Test as practice for this year's test. Scores from this test are used to award pay bonuses to teachers. The teachers worked at Wesley Elementary, which was also under investigation in 2003 when a former teacher accused school administrators of pressuring teachers to give test answers to students.
  (Tresaugue & Viren, 2006)

- 07/28/2006 Dallas Star-Telegram:
  The Texas Education Agency announces it will investigate testing irregularities at 609 schools from the 2005 administration of the Texas Assessment of Knowledge and Skills. Four types of irregularities were reported in Texas: patterns of similar responses, multiple marks on answer sheets, large score gains compared to previous years, and unusual response patterns. State-appointed monitors will oversee future test administrations.
  (Brock, 2006)

- 07/04/2006 Baltimore Examiner:
  Officials revoke the certificates of two fourth-grade teachers in Carroll County after they were accused of cheating on the Maryland School Assessments. One of the teachers admitted to copying questions from a previous test in order to create a practice worksheet for students.
  (Johnson, 2006)

- 06/30/2006 Brevard School District web site:
  Lori Backus, principal of Cocoa High School in Brevard, FL is accused of moving at least 54 9^{th} and 10^{th} grade special needs students into 11^{th} grade so that their FCAT scores would not count towards the school's grade (assigned by the state) in 2005 and 2006. As a result of an investigation into the allegations, Principal Backus was immediately removed as principal.
  (Brevard School District, 2006)

- 06/25/2006 Philadelphia Inquirer:
  Edison Schools fires Jayne Gibbs, principal at Parry Middle School in Chester, Pennsylvania for allegedly changing student test answers in 2005. Eighth graders at the school said the principal had given them the answers to questions on the Pennsylvania System of School Assessment. Gibbs is also accused of exempting special-education students from testing, violating state and federal rules. Edison Schools also asks the state and district to investigate exemplary test results at Showalter Middle School, where Gibbs served as principal from 2003-04.
  (Patrick & Eichel, 2006)

- 06/09/2006 Abilene Reporter-News:
  An elementary school in the Big Spring district in Texas is flagged for testing irregularities. Third-graders at Marcy Elementary were found to have too many erasure marks on the reading test in the 2005 Texas Assessment of Knowledge and Skills.

(Levesque, 2006)

- 05/23/2006 The Dallas Morning News:
  According to Caveon, a test security firm hired by the Texas Education Agency, almost 9% of schools had unusual scores on the Texas Assessment of Knowledge and Skills. Using statistical analyses, the firm found suspicious scores from 702 classrooms in 609 Texas schools in 2005. In one elementary school, 45 of the 262 students had identical answer sheets. An additional 29 students had perfect scores on the test. The chances of this happening naturally would be less than 1 in 1 trillion trillion trillion trillion trillion trillion (a 1 followed by 27 zeros).
  (Benton, 2006a)

- 05/21/2006 St. Louis Post-Dispatch:
  (excerpt) "Principal instructed teachers to encourage children to retry specific questions if the teachers thought the children knew the answer but had missed it on their first try."
  (CEA, 2007, p. 13)

- 04/17/2006 MSNBC:
  With permission from the federal government, nearly two million students' test scores are not counted when schools report progress by subgroups under the No Child Left Behind requirements. This is due to states being able to define the minimum number of students needed in a subgroup before scores are reported. In the past two years, almost half of all states have successfully petitioned the U.S. Department of Education to increase these minimums. An investigation concludes that about 1 out of every 14 test scores are not being counted under appropriate racial categories. The scores from more than 24,000 students in Missouri, 257,000 in Texas, and 400,000 in California are not being counted.
  (Associated Press, 2006b)

- 04/11/2006 The Columbus Dispatch:
  The Ohio Department of Education is investigating possible security breaches on the 2006 state tests. According to the department, 11 districts are investigating security breaches. The allegations include opening sealed boxes of test booklets early and teachers helping students cheat on the exams. Lora DeCarlo, a teacher at Franklin Middle School, was suspended without pay for 10 days. According to the teacher, she reviewed some student answer sheets and returned their tests to them with pages open to the items they needed to review. Other Ohio teachers accused of helping students cheat on tests in 2006 have resigned. Two years ago, a Hilliard teacher and a Reynoldsburg administrator resigned after acknowledging they broke test rules.
  (Richards, 2006b)

- 03/28/2006 The Baltimore Sun:
  (excerpt) "Teacher took notes based on the test administered last year and created worksheets for her pupils for this year's test. She also shared the worksheet with other teachers. Some of these other teachers, no knowing the origin of the questions on the worksheet, alerted the principal to similarities between the worksheets and this year's test."
  (CEA, 2007, p. 13)

- 03/08/2006 – 06/16/2006 Philadelphia Inquirer:
  Joseph Carruth, principal of Charles Brimm Medical Arts High in Camden, New Jersey, is fired after accusing Assistant Superintendent Luis Pagan of pressuring him to alter student answers on the 2005 High School Proficiency Exams. Carruth was allegedly told to create his own answer key and change answer sheets after the test was administered. The test scores from the high school significantly dropped the following year. The state also investigated two elementary schools for alleged cheating. Michael Mimms, principal of Sumner Elementary, is put on administrative leave after it is discovered that he possessed opened copies of the 2006 TerraNova exam and distributed it to teachers.

(Kummers & Burney, 2006abc)

- 02/07/2006 Memphis Eyewitness News:
  Teachers in Memphis schools are being investigated for test irregularities. According to the Tennessee Department of Education, an unusually high number of erasure marks were found on student exams. In many cases, incorrect answers were changed to correct answers.
  (Memphis Eyewitness News, 2006)

- 01/12/2006 New York Daily News:
  Fifth-grade students in Brooklyn were allegedly given actual copies of an exam to use as practice. Some students at Public School 58 in Cobble Hill reported that they recognized passages and questions from the test. Joyce Plus-Saly, the school principal, allegedly gave the materials to teachers to share with students, not knowing the questions would be used on the actual test.
  (Lucadamo, 2006)

- 12/23/2005 WCBS-TV New York:
  Ross Rosenfeld, a teacher at Junior High School 14 in Sheepshead Bay, was fired from his job after secretly recording conversations with the school principal. According to Rosenfeld, the recordings show that administrators ignored cheating on a state social studies exam. Rosenfeld was allegedly told to ignore a student who was found to have a cheat sheet during an exam.
  (Lyon, 2005)

- 09/29/2005 Pittsburgh Post-Gazette:
  Beth Boysza, a fourth-grade teacher in Pittsburgh, is suspended after being accused of helping her students on a math test in 2003. Boysza allegedly put Post-It® notes in the test booklets, providing students with special test instructions. She also is alleged to have re-read test questions to students. Boysza argues that she was simply providing accommodations to students, following directions provided by the district and test developer.
  (Ove, 2005)

- 09/19/2005 The Courier-Journal in Louisville, Kentucky:
  Following two cheating scandals, the Indiana Professional Standards Board increased the consequences for teachers who are caught helping their students cheat on tests. A teacher in Muncie, IN allegedly tapped her students on the shoulder to notify them of incorrect answers. A principal at Shakamak Elementary School in Jacksonville was found to have modified test questions and give them to students before the test administration. Both educators were caught after parents or state education officials noticed unusually large increases in school test scores.
  (Hupp, 2005)

- 08/29/2005 Union-Tribune in San Diego, CA:
  A teacher in Vista, CA was transferred to another school after allegations that she cheated on the California Standards Test. The teacher had allegedly put helpful materials on the classroom walls. Nearly half the students in the classroom reported that they had been told correct answers. The teacher was caught after a student reported the unusual behavior to her parents.
  (Jenkins, 2005)

- 06/28/2005 Free Republic:
  Isben Jeudy, a Long Island high school assistant principal, is arraigned after allegedly giving his son the answers to a history Regents exam. An official caught Jeudy's son with blue writing on his hand – the writing reportedly had answers to 35 of the exam questions.
  (Eltman, 2005)

- 05/24/2005 St. Louis Post-Dispatch:
  (excerpt) "Teachers prompted students with hand signals and pointed to answers."

(CEA, 2007, p. 13)

- 05/16/2005 Seattle Post Intelligence:
  Lisa Poitras alleges that her daughter's teachers at Lake Dolloff Elementary have cheated on exams for two consecutive years. The teachers allegedly check student answers, give assistance, and urge students to make corrections on the Washington Assessment of Student Learning. Poitras is reported to have said "her daughter was made to erase and rewrite her answer to a question so many times that she wore a hole through the booklet page and had to reinforce it with scotch tape."
  (Blanchard, 2005)

- 05/09/2005 Honolulu Advertiser:
  The Hawaii Department of Education is investigating reports of cheating on the Hawaii State Assessment. Eighth-grade students were allegedly given test questions and answers to prepare for the test administration. An anonymous school employee notified the newspaper that teachers were given review sheets with actual test items on them.
  (Shapiro, 2005)

- 05/04/2005 WHO TV in Des Moines, Iowa:
  Gene Zwiefel, a seventh-grade teacher in the Adel district, resigns after allegations were made that he quizzed students on materials found in the actual Iowa Tests of Basic Skills. According to David Frisbie, director of the Iowa Testing Programs, similar incidences have occurred at four other Iowa Schools.
  (WHO TV, 2005)

- 05/05/2005 Houston Chronicle:
  (excerpt) "Teachers signaled students by tapping them on their shoulders to let them know an answer was wrong."
  (CEA, 2007, p.13)

- 05/03/2005 Atlanta Journal-Constitution:
  Following an investigation of cheating in Texas, Georgia begins an investigation of its own test results. While no high-profile cheating case emerged in Georgia, 159 educators were sanctioned for test administration problems in the past five years.
  (Ghezzi, 2005)

- 05/03/2005 Star-Telegram in Texas:
  Two teachers at A.M. Pate Elementary School are no longer working after allegedly giving students answers to the Texas Assessment of Knowledge and Skills. One of the teachers, Georgia Johnson (a 25-year veteran), had 18 of the 19 students in her class pass the test. Six of her students had perfect scores. The other teacher, Mildred Lawrence-Medearis (17 years experience), had all 29 of her students pass the reading and math exams.
  (Garza, 2005)

- 04/13/2005 Rockford Register Star:
  The Illinois Department of Education is investigating Tiffany Parker, principal of Lewis Lemon Elementary School in Rockford, for allegedly altering student answers in 2003.
  (Watters, 2005)

- 04/30/2005 St. Louis Post-Dispatch:
  "After the allotted time for testing, a teacher told students to fill-in answers for questions they had left blank."
  (CEA, 2007, p. 13)

- 04/13/2005 NBC 6 in Miami, Florida:
  The Florida Department of Education has reassigned Nicholas Emmanuel, principal of West View Middle School, after he allegedly helped students cheat on the Florida Comprehensive Assessment Test.
  (NBC 6, 2005)

- 03/24/2005 Philadelphia Inquirer:
  Shirley Neeley, Pennsylvania State Education Commissioner, moves to dissolve the Wilmer-Hutchins Independent School District board after 22 educators were found to have cheated on the Texas Assessment of Knowledge and Skills. The teachers allegedly ordered students who finished the test early to fix answers on other students' answer sheets.
  (Mezzacappa et. al, 2005)

- 02/18/2005 The Ithaca Journal in Ithaca, New York:
  Robert Blair, a fourth grade teacher with 19 years experience at Palmer Elementary School, resigns after administrators discover altered answer sheets on his students' state English Language Arts tests. Based on an analysis of erasures, 17 or 18 of the 22 students in his class had their answer sheets altered. The report states that there were 14 proven cases of teacher cheating in 2003-04 in New York.
  (Associated Press, 2005)

- 01/31/2005 WRAL Raleigh-Durham, North Carolina:
  Following rumors of test misconduct at Sallie B. Howard School for the Arts and Education, North Carolina administrators report there have been at least 10 investigations into testing irregularities. In that time, two teachers had their licenses revoked and a third case is in litigation.
  (Carlson, 2005)

- 01/11/2005 Christian Science Monitor:
  The Houston Independent School District launches an investigation into suspicious results on the 2004 administration of the TAKS. Recent examples of reported educator cheating include: (a) a third grade teacher in Indiana was suspended for allegedly tapping students on the shoulder when they marked wrong answers, (b) a fifth grade teacher in Mississippi was fired for allegedly helping students on the writing portion of a test, (c) nine Arizona school districts discarded test results because teachers allegedly read the test to students and gave students extra time.
  (Axtman, 2005)

APPENDIX B: STATISTICAL DETECTION INDICES

Statistical methods to detect cheating do not, in fact, detect cheating. These methods, first developed in the 1920s to detect student cheating on multiple-choice tests, measure the likelihood of observing score gains, erasures, or answer patterns from student answer sheets. While the methods can identify unlikely large score gains, an improbably number of erasures, or unusual patterns of answers to items, they cannot determine if these events were due to cheating or simply due to chance.

Furthermore, the methods cannot identify all forms of cheating. They can only attempt to detect cheating due to educators giving answers to students or changing student answer sheets.

While most statistical detection methods were developed to detect student cheating, they can also be used to identify possible educator cheating. After all, if multiple students within a class or school are flagged as having unusual patterns of answers or erasures, the educator in charge of that class or school may have cheated.

Early Developments

Saupe (1960) summarizes the development of statistical detection methods to detect students who copy answers from other students. Bird (1927, 1929) developed three empirical approaches to detect possible copying in which the number of matching incorrect answers between two student tests is compared to the distribution of identical incorrect answers observed from a large random sample of answer sheet pairs (Saupe, 1960, p. 476). Because the number of incorrect answers depends upon the ability level of the student, the empirical distribution was based on random samples of test pairs from students with similar total scores to the suspected cheater. If the tests from the suspected cheater and the source student (from whom the cheater allegedly copied) were found to have an unusually large number of identical incorrect responses in comparison to this empirical distribution, then the suspected cheating could be verified.

In an application of his method, Bird describes a test administration in which the test proctor observed suspicious behaviors from four examinees. Bird calculated an average of 4.0 identical incorrect answers from a random sample of pairs of tests from examinees not suspected of cheating. The suspected cheaters had 17, 25, 28, and 31 identical incorrect answers on the 149-item test. As validation of his method, Bird reports that three of the suspected cheaters "confessed guilt when confronted with the evidence" (Bird, 1927, p.261).

From Empirical to Chance Models

Rather than taking the time to develop the empirical distribution, Dickenson (1945) developed a method to determine the likelihood of identical answers occurring by chance. This method simply compares the actual number of identical incorrect answers on a pair of answer sheets to the expected number based on the number of possible responses to each item. Under this method, it is assumed that each incorrect item response is equally likely to be chosen by students. If k is the number of possible responses to each item, then $(k-1)/k^2$ is the expected proportion of incorrect answers on one test that are identical to another test. Dickenson suggested that if the observed proportion of identical incorrect answers is more than twice the expected proportion, then copying is implied (Saupe, 1960, p. 476).

Anikeeff developed another chance model using the binomial distribution to determine the likelihood of observing a specific number of identical incorrect answers between two tests. The number of observed identical incorrect answers between a pair of tests is compared to a binomial distribution with a mean of N, and a standard deviation of $\sqrt{Np(1-p)}$, where N is the number of wrong responses by the suspected cheater and p is the reciprocal of the number of possible responses to each item (Saupe, 1960, p.

476). A low likelihood of observing that number of matching incorrect answers may indicate copying.

In an application of his method, Anikeef concludes that his method is not effective at detecting copying. He concluded that his method would be useful in situations in which an examinee copies more than 16% of the answers from another examinee (Anikeeff, 1954).

Bellezza and Bellezza (1989) developed a method similar to Anikeef's method called Error Similarity Analysis (ESA). This method, used by the *Scrutiny!* software package (Advanced Psychometrics, 1993), calculates the total number of times all pairs of examinees chose identical incorrect answers for each item. The probability of observing a given number of identical incorrect answers is estimated by the binomial distribution:

$$\frac{w!}{c!(w-c)!}P^{c}(1-P)^{w-c}$$

where c is the number of common items answered incorrectly by a pair of examinees, w is the number of items for which the pair of examinees had identical incorrect responses, and P is the estimated probability of two examinees selecting an identical incorrect answer. Using this equation, or a method based on a standard normal approximation, Bellezza and Bellezza were able to determine the likelihood of observing a specific number of identical incorrect answers between two examinees.

Holland (1996) describes another popular method to detect possible cheaters called the K-Index. This index, used by the Educational Testing Service (ETS), may be the most popular method currently used (Cizek, 1999, p. 141). Although limited information about this index exists, Holland describes it as a method used to "assess the degree of unusual agreement between the incorrect multiple-choice answers of two examinees" based on an estimate of the probability two examinees would agree on a response by chance (Holland, 1996, p. 5). The index uses the binomial distribution to

model this probability. Sotaridona (2001) developed two indices, S1 and S2, similar to the K-Index except using the Poisson distribution to model the probability of observing a specific number of identical examinee responses.

Incorporating More Information

Acknowledging the limitation in methods that only analyze matching incorrect answers, Saupe (1960) developed his method to detect copying on multiple-choice tests. In this method, the total number of items on the test, K, is partitioned into $K = R_i + R_j - R_{ij} + W_{ij}$, where R_i and R_j are the number of correct responses for students i and j, respectively; R_{ij} is the number of items both students answered correctly; W_{ij} is the number of items both students answer incorrectly; and w_{ij} is the number of items in which both students gave matching incorrect answers.

Under chance conditions, the expected number of items answered correctly by both students would be the proportion of all items answered correctly by student i multiplied by the number of items answered correctly by student j: $ER_{ij} = \frac{1}{k} R_i R_j$. Thus, the regression of R_{ij} on the product $R_i R_j$ is of interest. This regression line can be written as: $\hat{R}_{ij} = b_{r1} R_i R_j + b_{r0}$.

The distance of an observed point $\left(R_i R_j, R_{ij}\right)$ from the regression line can be used to evaluate the observed degree of correspondence between the items answered correctly by a pair of students. If this distance exceeds ts_r, where t is the appropriate value from the t-distribution and s_r is an appropriate estimate of the standard error of estimate of R_{ij}, then the assumption of chance correspondence can be rejected at a specified confidence level (assuming a bivariate normal distribution of R_{ij} and $R_i R_j$). A correspondence index can be written as:

$$CI = \frac{R_{ij} - b_{r1} R_i R_j - b_{r0}}{ts_r}.$$

A correspondence index greater than 1.00 is equivalent to rejecting the null hypothesis of chance correspondence between the items answered correctly by a pair of students.

The same logic is used to determine the correspondence between the incorrect answers from two students. If each item has k possible responses, the expected number of matching incorrect answers due to chance is:

$$Ew_{ij} = \frac{1}{k-1} W_{ij} .$$

Using the regression of w_{ij} on W_{ij}, the correspondence index would be:

$$CI = \frac{w_{ij} - b_{w1} W_{ij} - b_{w0}}{ts_w}$$

Saupe suggests an advantage to analyzing the correspondence of correct and incorrect answers separately is that the evidence provided by both indices is non-overlapping and, therefore, complementary. In applying his model to a random sample of 150 pairs of tests, Saupe's correspondence indices identified 6 suspicious pairs. In an attempt to validate the results, Saupe examined seating charts and discovered that 5 of the 6 suspicious pairs came from students in adjacent seats. Saupe admits the main disadvantage of his method is its use of a chance model – it is not reasonable to assume students randomly answer test questions (Saupe, 1960).

Attempting to overcome this disadvantage, Angoff (1974) developed 8 more indices to detect examinees copying on tests. Angoff's methods were all based on developing distributions of identical responses made by pairs of non-cheating examinees. The methods only differ in the combinations of independent and dependent variables used to develop the bivariate distributions. The degree to which an examinee's observed value on the dependent variable, conditioned on the observed value of the independent variable, deviates from the mean of the dependent variable from the distribution provides an index of cheating.

Angoff found that six of his indices were not effective in detecting cheating. Of the remaining indices, Angoff favored the method called the B Index. To use this index, the bivariate distribution of W_iW_j and Q_{ij} is estimated from all examinees, where W_iW_j is the product of the number of incorrect answers from two examinees and Q_{ij} is the number of identical incorrect answers for both examinees. For a pair of examinees, A and B, the observed values W_aW_b and Q_{ab} are calculated. The following test statistic can then be used to determine whether the observed value of Q_{ab} is significantly different from the mean value of Q_{ij}:

$$t = \frac{Q_{ab} - \bar{Q}_{ij}}{S_{Q_{ij}\ W_iW_j}}.$$

While Saupe and Angoff used information from both incorrect and correct responses, Frary (1977) developed two indices based on estimating the probability of an examinee choosing a correct response, choosing each incorrect response, or choosing to omit each item. After dismissing his first index, Frary developed the following formula for his g_2 index:

$$g_2 = \frac{C - \sum_i \hat{P}(k_{ia} = k_{ib})}{\sqrt{\sum_i \hat{P}(k_{ia} = k_{ib})\left[1 - \sum_i \hat{P}(k_{ia} = k_{ib})\right]}},$$

where C is the number of identical answers for a pair of examinees and $\hat{P}(k_{ia} = k_{ib})$ is the probability that an examinee would choose the identical response of another examinee. Frary used piecewise linear functions of total test scores to estimate this probability.

After applying his method to actual test data and recommending its use to prevent cheating, Frary acknowledged three limitations. First, in order to use his method, one examinee must be identified as "the copier" and another examinee must be identified as "the source." This will not always be practical in large-scale testing situations. Second,

the g_2 index assumes that the probabilities of an examinee choosing each response to an item are constant, regardless of examinee ability. Third, Frary found that his method decreased in effectiveness for easier tests, stating, "If no examinees can answer as many as 90% correctly, the potential for detection is greatly enhanced" (Frary, 1977, p.253).

Hanson and Brennan (1987) continued to compare responses between pairs of examinees in their development of two more indices to detect possibly copying. The first method, Pair 1, uses the number of identical incorrect responses between a pair of examinees along with the length of the longest string of identical responses. The second method, Pair 2, uses the same information along with the percentage of maximum possible identical incorrect responses between two examinees.

In comparing their methods to the methods developed by Angoff (1974) and Frary (1977) on a simulated data set, Hanson and Brennan conclude that "it might not make a great deal of difference which of the statistical methods of investigating copying considered here are used" (p. 21). They do, however, recommend their method based on the interpretability of their indices.

Controlling for False Positives

In evaluating the effectiveness of the previously developed indices, Post (1994) concludes that while the indices may be used to scan for potential cheaters, "many existing statistical tests designed to detect copying on multiple-choice exams understate the Type I [false positive] error" (p. 140). Because this Type 1 error may be higher than specified, Post discourages using the indices to make accusations of cheating. Post attributes this inflated Type I error rate to the difficulty in estimating the probability of an examinee choosing each possible response to an item.

In an attempt to improve the estimation of item response probabilities and reduce the Type I error rate, Wesolowsky (2000) made a slight modification to Frary's method. Whereas Frary used piecewise linear functions of raw scores to estimate probabilities,

Wesolowsky uses smooth distance iso-contours from location theory for estimation (p. 912). Also, while previous methods made assumptions about which examinee copied from another, Wesolowky's method simply examines the number of matching answers and ignores other suspicious patterns such as strings of identical answers. In developing a computer program to analyze answer sheets and employing a Bonferroni adjustment to control for overall Type I error rate, Wesolosky recommends his method as an effective way to screen for potential cheaters. This method, used in 2007 to scan for cheaters on the Texas Assessment of Knowledge and Skills, flagged more than 50,000 examinees as potentially having cheated (Benton & Hacker, 2007a, 2007b).

Incorporating Item Response Theory

Other researchers improved the estimation of the probability of an examinee choosing each possible response to an item by developing indices based on item response theory (IRT) models. In IRT models, the probability of an examinee choosing each response to an item is a function of the examinee's latent ability, θ, and characteristics of each possible item response. The item response characteristics of interest depend on the IRT model being used.

For the three-parameter logistic model, the probability of examinee a correctly answering dichotomously scored item i can be expressed as:

$$P_{ia} = P_{ia}(\theta_a) = P_{ia}(u_i = 1 | \theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7 a_i (\theta_a - b_i)}},$$

where

a_i = the item discrimination parameter,

b_i = the item difficulty parameter,

c_i = the guessing parameter,

θ_a = the latent ability of examinee a,

and μ_i = the examinee's scored response to the item.

Using maximum likelihood or Bayesian methods, values for the item response and examinee ability parameters can be estimated under the assumptions of the specified IRT model. Under the assumption of local independence, the probability of observing a string of n item responses from an examinee with ability θ_a is equal to the product of the probabilities of the individual item responses:

$$P(u_1, u_2, ..., u_n | \theta) = \prod_{i=1}^{n} P_i^{u_i} (1 - P_i)^{(1-u_i)}.$$

Thus, given an observed string of responses from an examinee, the above formula can be used to estimate the likelihood of observing that response string or the probability of observing a different string of responses.

Rather than simply estimating the probability of an examinee answering dichotomously scored items correctly or incorrectly, IRT models can be used to estimate the probability of an examinee choosing each possible choice to a multiple-choice item. Bock's Nominal Model calculates the probability of choosing response u on multiple-choice item g with m possible responses as:

$$P_u = \frac{\pi_u}{\sum_{h=1}^{m} \pi_h},$$

where π_h represents a scale value directly related to the probability that response h is chosen on a specific test item. This model can be reparameterized as:

$$P_u(\theta) = \frac{e^{(a_u \theta + c_u)}}{\sum_{h=1}^{m} e^{(a_h \theta + c_h)}},$$

from which item response discrimination and difficulty parameters, a and c, and examinee ability parameter θ can be estimated. Again, under the local independence assumption, the probability of observing a specific string of item responses can be estimated from these item response and examinee ability estimates.

Person-Fit and Aberrant Response Indices

The application of IRT models to detect possible cheating has been theorized via person-fit and aberrant response indices. These indices measure the extent to which the observed pattern of responses from an examinee with ability level θ deviates from the response pattern expected under the chosen IRT model. For example, an examinee whose ability exceeds the difficulty level (b) of an item would have a high probability of answering that item correctly. Likewise, an examinee whose ability is less than the difficulty of an item would have a high probability of answering that item incorrectly. When an examinee's response string fits this pattern across most or all items on the test, the model "fits" the person. Aberrant response strings (high-ability examinees incorrectly answering easy items, low-ability examinees correctly answering difficult items, or examinees choosing unusual responses on a multiple-choice test) indicate poor model fit. Person-fit and aberrant response indices measure the degree to which the chosen IRT model fits the observed responses from an examinee.

More than fifty person-fit indices have been developed to detect aberrant responders (Karabatsos, 2003; Meijer & Sijtsma, 2001; Thiessen, 2004). These indices, displayed in Figure B1, attempt to detect students who provide unusual responses due to luck, language deficiencies, random guessing, low-motivation, misaligned answer sheets, or cheating (Meijer, 1996).

Person-fit indices can be classified into three categories: deviation-based, covariance-based, and likelihood-based. Deviation-based indices, such as Wright and Stone's (1979) Outfit Mean Square index, sum the squared standardized differences between an examinee's scored response to an item and the expected probability of that correct response. A large difference would indicate a disagreement between the model and the examinee.

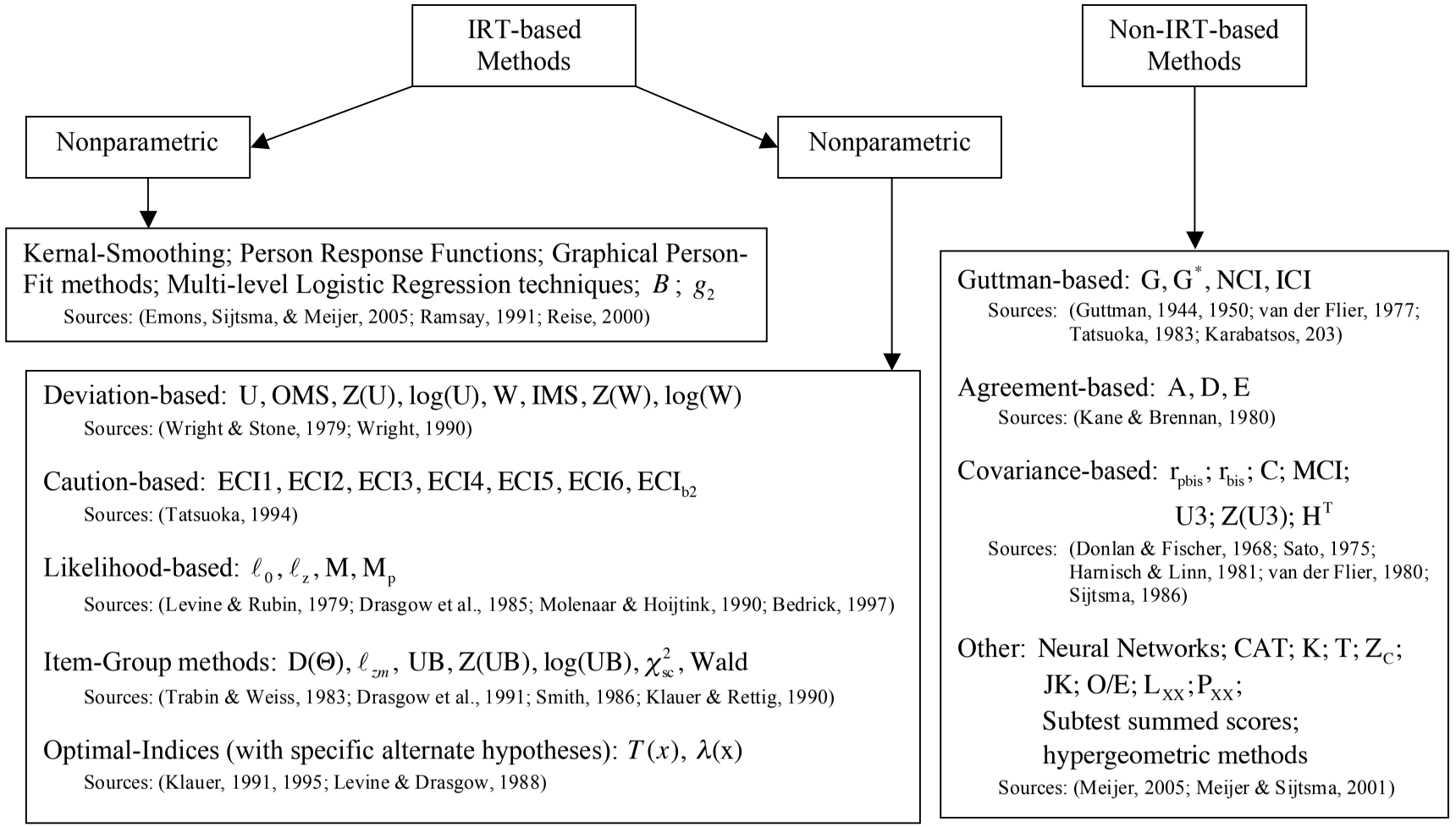


Figure B1 Aberrant Response Detection Methods and Indices

Note: See (Karabatsos, 2003; Meijer, 1996; Meijer & Sijtsma, 2001; Thiessen, 2004) for formulas and discussion of these methods.

Covariance-based indices, such as Tatsuoka's (1984) Extended Caution Indices, measure the degree to which an examinee's item responses deviate from the Guttman Perfect Pattern. Specifically, these indices calculate the ratio of the covariance between an examinee's responses and item difficulty estimates to the covariance between the average probability of correct responses across all examinees (estimated via IRT models) and the item difficulty estimates.

Likelihood-based indices, such as Levine and Rubin's (1979) log-likelihood function ℓ_0 evaluate the shape of a likelihood function. Given an examinee's responses to a set of test items with known item parameters (a, b, c), the ability level of an examinee, θ_a, can be estimated by maximizing the likelihood function:

$$\ell_0 = \ln\left[L(u|\theta;a,b,c)\right] = \ln\left[\prod_{i=1}^{n} P_i^{u_i} Q_i^{(1-u_i)}\right] = \sum_{i=1}^{n}\left[u_i \ln(P_i) + (1-u_i)\ln(Q_i)\right].$$

The value of θ_a that maximizes this function represents the examinee's most likely ability level given the observed item responses and the estimated item parameters. Examinees whose item responses conform to the IRT model produce likelihood functions with relatively high maximum values. Examinees whose responses deviate from what is predicted by the IRT model produce low maximum values of the likelihood function (Davey, et. al, 2003, p.6). Thus, the relative magnitude of ℓ_0 can be used to identify examinees whose responses are aberrant.

Aberrant Response Indices to Detect Examinee Cheating

While person-fit indices detect many forms of aberrance, some indices were developed specifically to detect examinee cheating. The first set of these indices is referred to as optimal person-fit statistics. Levine and Drasgow's (1988) $\lambda(x)$ index was developed to test the null hypothesis of normal examinee responses (based on a chosen IRT model) against the alternative hypothesis that an examinee's responses are consistent with a specific aberrant response model. Thus, if a researcher can specify and estimate

an alternative hypothesis of examinee cheating, optimal person-fit statistics can provide the maximum detection rate for aberrance (Karabatsos, 2003, p. 282).

Unfortunately, studies have shown that aberrant response indices are ineffective at detecting cheating (Chason & Maller, 1996; Iwamoto, Nungester, & Luecht, 1996). Demonstrating this ineffectiveness, test security firm Caveon analyzed a simulated data set using their six aberrant response indices. With the data set simulated so that examinees cheat on 10% of the test items, the firm's six aberrant response indices were only able to detect 41 (1.2%) of the 3,283 simulated cheaters and none of the five simulated cheating schools (Impara, et al., 2005).

Wollack argues that IRT-based aberrant response indices are inadequate in detecting cheating (specifically, student copying) because the statistical significance of these indices "does not depend on the similarity between the suspected copier's responses and those of a neighboring examinee" (p. 307). He stated that previous attempts to detect cheating, based on Classical Test Theory (CTT), are also inadequate. Wollack argues that since CTT item statistics are dependent on the sample of examinees tested, the expected similarity between a pair of examinees depends "largely on the performance of the other examinees in the sample, rather than only on the two examinees of interest" (p. 307). In order to overcome the apparent limitations of CTT-based indices and IRT-based aberrant response indices, Wollack developed his ω-index to specifically detect copying.

The ω-index is similar to Frary's g_2 index in that it attempts to estimate the probabilities associated with each possible item response. It differs from the g_2 index in that it uses Bock's Nominal Model to model these probabilities. The ω-index also attempts to control for Type I errors by examining the seating chart used in test administration and analyzing only pairs of examinees who sit physically close enough to make copying possible.

In applying the ω-index, each examinee is analyzed as a possible copier. If the test is administered in a classroom with a rectangular seating chart, the examinees sitting

to the copier's left, right, front left, front center, or front right are analyzed as potential sources for copying. Assuming item responses are locally independent, as is necessary to use IRT models, the expected number of identical responses and its associated variance can be modeled with the binomial distribution. Thus, an index similar to the g2 can be compared to the standard normal distribution to evaluate the statistical significance of the degree of similarity between two examinees' item responses:

$$\omega = \frac{C - \sum_i \hat{P}(k_{ia} = k_{ib} | \theta_a, K_b, \xi)}{\sqrt{\sum_i \hat{P}(k_{ia} = k_{ib} | \theta_a, K_b, \xi)\left[1 - \sum_i \hat{P}(k_{ia} = k_{ib} | \theta_a, K_b, \xi)\right]}},$$

where C is the number of identical answers for a pair of examinees and $\sum_i \hat{P}(k_{ia} = k_{ib} | \theta_a, K_b, \xi)$ is the probability that examinee a would choose the identical response of examinee b given the ability θ_a of examinee a, the item responses K_b of examinee b, and the matrix of item parameters, ξ.

Wollack (1997) compared his index to Frary's g_2 index on two data sets with three types of simulated copiers. The first type of copying was random copying, in which randomly selected items were simulated as being copied from another examinee. The second type was difficulty-weighted copying, in which the more difficult items were simulated as being copied. The third type was random-strings copying, in which strings of 4 consecutive items were simulated as being copied. Between 10- 40% of items were simulated as being copied by 5% of examinees for each type of copying. Wollack ensured that in this simulation, simulated copiers copied answers from examinees with higher ability estimates sitting close to them, as would be expected in reality. Based on these simulations, Wollack concludes that the ω-index is more effective in both detecting copying and controlling Type 1 error rates than the g_2 index when a seating chart is available (Wollack, 1997, p.312).

Adjacent Seating Methods

Similar to Wollack's ω-index, Kvam (1996), Roberts (1987), and the National Board of Medical Examiners (Cizek, 1999) developed methods that require knowledge about which examinees sat adjacent to others. In applying Kvam's method, examinees in a classroom are randomly administered two forms of the exam. Minor changes are made to the questions so that the two multiple-choice forms have different answers. After administering the exams, maximum likelihood methods are used to estimate the probability that an examinee copies an item response from an adjacently-seated source given that the examinee does not formulate an answer to the question (Kvam, 1996, p.239).

The score-difference method developed by Roberts (1987) also relies on two test forms randomly administered to examinees. After administering the test forms, each answer sheet is scored using the answer keys from both test forms. The difference between the scores obtained from the two answer keys provides an indication of possible cheating. While conceptually simple, Roberts concludes that his method is "seriously flawed and has little to recommend" (p. 77).

In administering its exams, the National Board of Medical Examiners (NBME) uses two methods to detect potential cheating. The first method, adjacent-nonadjacent analysis, requires the test to consist of at least two parallel parts. Each part must be administered in separate testing sessions in which test takers are randomly assigned to seat locations. Under these conditions, it is expected that the response patterns from any two examinees would have the same degree of similarity regardless of where the examinees are seated. If two examinees are seated adjacently during at least one test session and it is discovered that their responses are more similar in those sessions than would be expected if they were seated apart, then potential copying has been detected.

The NBME method uses a simple 2x2 chi-square test for independence to detect potential cheaters. The test statistic is calculated from Table B.1 using the formula given below, and then evaluated for significance:

$$\chi_1^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}.$$

Table B.1: Results from 2005 state and NAEP tests of 8th grade mathematics

	Number of items answered incorrectly		
Seating Location	Identical Responses	Different Responses	Total
Adjacent	a	b	$a+b$
Nonadjacent	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

Methods to Detect Educator Cheating

The previously discussed methods all attempt to detect cheating (usually in the form of copying) by students on tests. Only two serious attempts have been made to detect potential educator cheating on tests. The first method, erasure analysis, is currently used by several states in auditing their testing programs. The second method, developed by Jacob and Levitt (2003), was used to detect cheating educators in Chicago Public Schools.

Some erasure analysis methods attempt to detect cheating by identifying answer sheets with an unusual number or pattern of erasures. It is assumed that a large number of erased answers might indicate an educator who is manipulating answer sheets. Other erasure analysis methods only look at wrong-to-right erasures – items in which an incorrect answer was erased and changed to a correct answer.

The Louisiana Educational Assessment Program has written and implemented erasure analysis procedures to audit its state testing program. These procedures require

the test scoring contractor to scan every answer sheet for wrong-to-right erasures and compute the mean and standard deviation for each subject at each grade level. The scoring contractor must then flag examinees whose wrong-to-right erasures exceed the state mean by more than four standard deviations as potential cheaters. The number of erasures for each potential cheater, along with the proportion of wrong-to-right erasures found on their answer sheet, are then reported to the State Superintendent of Education and the scores from the potential cheaters are voided from official records (Louisiana Educational Assessment Program, 2003).

One problem with using erasure analysis to flag potential cheaters is determining how many erasures indicate potential cheating. To address this problem, Qualls (2001) examined 4,553 answer sheets from the 1995-96 administration of the Iowa Tests of Basic Skills to determine the typical erasure behavior of examinees in a low-stakes testing environment. Qualls found that more than 90% of examinees changed 3 or fewer answers per test and only 2% of examinees changed more than 6 responses. While the erasure behavior depended on the test subject, less than 7% of item responses were erased, on average. Furthermore, the results indicate that it would be rare for an examinee to erase and change more than 15% of items on a single test. Qualls found that the first or second items on the test were most frequently erased and that students with one erasure had about a 50% chance of changing an incorrect response to a correct response. In a focus on wrong-to-right erasures, Qualls found that examinees gained an average of between 0.167 and 0.494 points per erasure on the tests. Perhaps these results could be used to develop new and improve upon current erasure analysis methods for high-stakes tests.

The other attempt to detect educator cheating on large-scale standardized tests was developed by Jacob and Levitt (2003). Jacob and Levitt set out to create a statistical index to identify educators, specifically teachers, who manipulate answer sheets.

The method used by Jacob & Levitt to detect potential cheating teachers is actually a combination of two indicators: (1) unexpected test score fluctuations and (2) unexpected patterns in student answers. Rather than focus on individual students, these indices are calculated at the classroom level.

Index #1: Unusual Test Score Fluctuations

On a vertically scaled test like the ITBS, the expected gain in student test scores from year-to-year can be estimated. The test score gains depend on a variety of factors (student ability, curriculum, teacher quality, etc.), but most students will experience growth in achievement each year. An unexpected test score fluctuation would occur when many students in a classroom experience large score gains one year followed by small gains (or even negative growth) the next year. To calculate this unexpected test score fluctuation index, Jacob & Levitt do the following:

- If the interest is in determining if a teacher manipulated answer sheets during year *t*, then test scores must be collected for year *t*, the previous year (*t-1*), and the next year (*t+1*).
- Find the average test score gains (in grade equivalent units) for all students in each classroom (the gain from year *t-1* to year *t* along with the gain from year *t* to year *t+1*.
- Find the percentile rank of each classroom's average test score gains relative to all other classrooms in that same subject, grade, and year. The percentile rank of growth from year t-1 to year t will be called rank gain_t while the rank of growth from year *t* to year *t+1* will be called rank gain_{t+1}.
- The index is calculated as: $\text{Index \#1} = \left(\text{rank gain}_t\right)^2 + \left(1\text{-rank gain}_{t+1}\right)$. As the formula shows, classrooms with relatively large score gains followed by low gains the next year will yield large values of this index. The percentile ranks are

squared in computing the index to give relatively more weight to big score gains and big score declines.

- Teachers whose classrooms yield values in the top 95th percentile of this index are identified as having unusual test score fluctuations.

Table B.2 shows an example of this index. The test score gains for three classrooms are displayed. Classroom 1 represents an average classroom, growing from 2.9 to 4.1 to 5.2 grade equivalent units over the course of two years. Classroom 2 represents a classroom with an excellent teacher (during year t). This excellent teacher was able to drastically increase test scores from year *t-1* to year *t*. The teacher of Classroom 3 also was able to get drastic increases in test scores. The difference between Classrooms 2 and 3 can be seen by the changes in test scores once the students moved on to the next grade (with a different teacher). Whereas the students from Classroom 2 were able to build upon their score gains (as we would expect from students who were taught by an excellent teacher), the students in Classroom 3 experienced test score declines. This decline in test scores may indicate that the gains experienced the previous year did not represent genuine gains in achievement. Thus Classroom 3 earns a high score on Index #1 and the teacher of that classroom is identified as a potential cheater.

Jacob & Levitt experienced with other measures of unusual test score fluctuations (such as regressing test score gains on previous gains and student demographics). They concluded that these other, more complicated measures yielded information similar to their simple index.

Table B.2: Example data for Index #1

Year	Avg. Classroom Grade Equivalent			Change in Grade Equivalent Units		Index #1 (Percentile Rank)
	t-1	*t*	*t+1*	From *t-1* to *t* (percentile rank)	From *t* to *t+1* (percentile rank)	
Classroom 1	2.9	4.1	5.2	1.2 (59)	1.1 (56)	0.5400 (20)
Classroom 2	3.4	5.5	6.8	2.1 (92)	1.3 (77)	0.8993 (70)
*Classroom 3	3.1	5.2	4.7	2.1 (92)	-0.5 (1)	1.8265 (99*)
Avg. Classroom	3.3	4.4	5.3	1.1	0.9	0.5000 (50)

Index #2: Unexpected Patterns in Student Answers

Unusual fluctuations in test scores do not prove that a teacher manipulated student answer sheets. In fact, since the 95th percentile is used as a cut-off, Index #1 will always identify 5% of the classrooms as having unusual fluctuations. To determine which of these classrooms cheated (and which just experienced improbable fluctuations), Jacob & Levitt developed a second index to identify unexpected patterns in student answers.

The logic is this: The quickest way for a teacher to cheat is to alter the same block of consecutive items for students in the class (or instruct students in the classroom to change their answers to the same set of items). Thus, if a classroom experiences unusual test score fluctuations and the students in the classroom have unusual answer patterns (identical answers to the same block of items or unexpected correct answers to difficult items), then we have more reason to believe the teacher cheated.

To identify unexpected answer patterns, the researchers combine four measures of suspicious answer strings to calculate Index #2. These four measures will be briefly discussed.

The first measure focuses on identifying the most unlikely block of identical answers given by students on consecutive items. Using a multinomial logit model, the likelihood of each student choosing each possible answer on every item is calculated. This likelihood is based on the student's past test scores, future test scores, and demographic (gender, race, etc). All combinations of students and consecutive items are searched to find the block of identical answers that were least likely to have arisen by chance (controlling for classroom size).

Given student s in classroom c with answer j on item i, the model is:

$$P(Y_{isc} = j) = \frac{e^{\beta_j X_s}}{\sum_{j=1}^{J} e^{\beta_j X_s}},$$

Where X represents the vector of past test scores, future test scores, and demographics. The likelihood of a student's answer string for item m to item n is calculated as:

$$P_{sc}^{mn} = \prod_{i=m}^{n} P_{isc} .$$

This likelihood is multiplied across students in the class with identical responses in the string:

$$\tilde{P}_{sc}^{mn} = \prod_{\text{Students with identical strings}} P_{sc}^{mn} .$$

If each student in the classroom has unique responses from item m to item n, then there will be a distinct value of this index for each student in the class. If all students in the classroom have identical responses across these items, then there will only be one value of this index (and the value will be extremely small). The calculations are repeated for all strings from a length of 3 items to a length of 7 items.

Notice that the values of yielded by these calculations will be smaller as: (1) the number of students with identical responses increase, (2) the length of the string of identical responses increase. Thus, smaller values are associated with more improbable answer strings within a classroom.

The minimum value for each classroom is recorded as measure #1:

$$\text{Measure \#1} = \min_s(\tilde{P}_{sc}^{mn}).$$

The second measure calculates the degree of correlation in student responses across the test, especially for unexpected answers. The logic is that teachers who cheat will have students with highly correlated answers. To calculate this measure, the residuals for each item choice are calculated:

$$e_{ijsc} = \begin{cases} 0 - P(Y_{isc}), & \text{for the unchosen options} \\ 1 - P(Y_{isc}), & \text{for the chosen answer} \end{cases}.$$

Then, the residuals for each option are summed across students within the classroom:

$$e_{jc} = \sum e_{ijsc}.$$

The option residuals for the classroom are then summed for each item. At the same time, the residuals are (1) squared to accentuate outliers and (2) divided by the number of students in the class to normalize for class size (n):

$$v_{ic} = \frac{\sum_j e_{jic}^2}{n}.$$

Measure #2 is simply the average of these item residual values:

$$\text{Measure \#2} = \bar{v} = \frac{\sum_i v_{ic}}{n}$$

Higher values indicate classrooms with highly correlated answers.

The third measure calculates the variance in the degree of correlation across test items. With Measure #2, we might expect high correlations among student answers if a teacher emphasizes certain topics during the school year. If a teacher cheats by changing answers for multiple students on selected questions, the within-class correlation on those particular questions will be extremely high, while the degree of within-class correlation on other questions is likely to be typical. Thus, a teacher who changes answers on selected items will have a classroom with a large degree of variance in the correlation of responses across items.

This measure is calculated as the variance of item residuals from Measure #2:

$$\text{Measure \#3} = \sigma_{\text{v}} = \frac{\sum_{i}(v_{ic} - v_{c})^2}{ni}.$$

The fourth measure compares the answers of students within a classroom to the answers from other equally-able students in the sample. This measure can then detect students who miss easy items while answering difficult items correctly. Students whose answers follow this pattern may have had their answers influenced by a cheating teacher.

To calculate this measure, students are grouped by their total number correct scores on the test. Let A_s represent a specific total correct score. Let $q_{ic} = 1$ if a particular student answers item i correctly and zero otherwise. Then determine the proportion of students with total score A_s who answered each item correctly (call this quantity $\overline{q}_A$).

The deviations between a student's item score and the expected item score (based on equally-abled students) are squared and summed across items:

$$Z_{sc} = \sum (q_{isc} - \overline{q}_A)^2.$$

This deviation between this Z-value for each student and the average Z-value for all equally-abled students is then summed for all students within a classroom:

$$\text{Measure \#4} = \sum (Z_{sc} - \bar{Z}_A)$$

High values of this index indicate the answers from a large number of students in the classroom deviated from equally-abled students in other classrooms.

After completing the calculations, the classrooms are ranked on each of the four measures. The percentile ranks for each classroom on each measure are then combined to form the second index:

$$\text{Index \#2} = (\text{Measure 1 rank})^2 + (\text{Measure 2 rank})^2 + (\text{Measure 3 rank})^2 + (\text{Measure 4 rank})^2$$

Classrooms falling above the 95th percentile on this index are identified as having unusual answer patterns.

Combining Indices to Detect Cheating Classrooms

Jacob & Levitt argued that taken individually, the above two indices do not detect teachers who manipulate answer sheets. After all, there are always going to be (innocent) classrooms with unexpected score fluctuations and there are going to be (innocent) classrooms with improbable answer patterns. The key is to identify classrooms that yield high values on both indices.

In non-cheating classrooms, there is no reason to believe that the two indices would have a strong correlation. If a teacher manipulates student answer sheets, we would expect a strong correlation between the two indices. Therefore, educators whose classrooms appear above the 95th percentile on both indices are identified as potential cheaters.

APPENDIX C: NATIONAL TESTING CODES AND STANDARDS

Source: *The Standards for Educational and Psychological Testing* (AERA et al., 1999)

Developer(s): American Educational Research Association, American Psychological Association, National Council on Measurement in Education

Statements related to inappropriate testing behaviors:

Validity, the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests, is the most fundamental consideration in developing and evaluating tests (p. 9).

The usefulness and interpretability of test scores require that a test be administered according to the developer's instructions. Maintaining test security also helps to ensure that no one has an unfair advantage (p. 61).

5.1 – Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer
5.2 – Modifications or disruptions of standardized test administration procedures or scoring should be documented.
5.6 – Reasonable efforts should be made to assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means.
5.7 – Test users have the responsibility of protecting the security of test materials at all times (pp. 63-64).

Fairness requires all examinees to be given a comparable opportunity to demonstrate their standing on the construct(s) the test is intended to measure. Just treatment also includes such factors as appropriate testing conditions and equal opportunity to become familiar with the test format, practice materials, and so forth. Fairness also requires that all examinees be afforded appropriate testing conditions. Careful standardization of tests and administration conditions generally helps to assure that examinees have comparable opportunity to demonstrate the abilities or attributes to be measured (pp. 74-75).

Ideally, examinees would also be afforded equal opportunity to prepare for a test. Examinees should in any case be afforded equal access to materials provided by the testing organization and sponsor which describe the test content and purpose and offer specific familiarization and preparation for test taking. In addition to assuring equity in access to accepted resources for test preparation, this principle covers test security for nondisclosed tests (p. 75).

7.12 – The testing or assessment process should be carried out so that test takers receive comparable and equitable treatment during all phases of the testing or assessment process (p. 84)
11.7 – Test users have the responsibility to protect the security of tests, to the extent that developers enjoin users to do so
11.8 – Test users have the responsibility to respect test copyrights (p. 115)
13.11 – In educational settings, test users should ensure that any test preparation activities and materials provided to students will not adversely affect the validity of test score inferences.
13.12 – In educational settings, those who supervise others in test [administration] should have received education and training in testing necessary... (p. 148).
15.9 – The integrity of test results should be maintained by eliminating practices designed to raise test scores without improving performance on the construct or domain measured by the test (p. 168).

Source: *Code of Professional Responsibilities in Educational Measurement*, (Schmeiser et al., 1995)

Developer(s): National Council on Measurement in Education, National Association of Test Directors

Statements related to inappropriate testing behaviors:

Those who prepare individuals to take assessments and those who are directly or indirectly involved in the administration of assessments as part of the educational process, including teachers, administrators, and assessment personnel, have an important role in making sure that the assessments are administered in a fair and accurate manner. Persons who prepare others for, and those who administer, assessments have a professional responsibility to:

4.3: Take appropriate security precautions before, during, and after the administration of the assessment.
4.4: Understand the procedures needed to administer the assessment prior to administration
4.5: Administer standardized assessments according to prescribed procedures and conditions and notify appropriate persons if any nonstandard or delimiting conditions occur
4.6: Not exclude any eligible student from the assessment
4.7: Avoid any conditions in the conduct of the assessment that might invalidate the results
4.11: Avoid actions or conditions that would permit or encourage individuals or groups to receive scores that misrepresent their actual levels of attainment

Conducting research on or about assessments or educational programs is a key activity in helping to improve the understanding and use of assessments and educational programs. Persons who engage in the evaluation of educational programs or conduct research on assessments have a professional responsibility to:

8.3: Preserve the security of all assessments throughout the research process as appropriate

Source: *Code of Fair Testing Practices in Education*, (JCTP, 2004)

Developer(s): Joint Committee on Testing Practices, American Psychological Association, National Council on Measurement in Education, American Counseling Association, American Educational Research Association, American Speech-Language-Hearing Association, National Association of School Psychologists, National Association of Test Directors

Statements related to inappropriate testing behaviors:

Fairness implies that every test taker has the opportunity to prepare for the test (p. 2).

Test users should administer and score tests correctly and fairly.
B-1: Follow established procedures for administering tests in a standardized manner.
B-3: Provide test takers with an opportunity to become familiar with test question formats and any materials or equipment that may be used during testing (contradicts Popham)
B-4: Protect the security of test materials, including respecting copyrights and eliminating opportunities for test takers to obtain scores by fraudulent means
D-1: Inform test takers in advance of the test administration about the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. Make such information available to all test takers

Source: *Standards for Teacher Competence in Educational Assessment of Students*, (NEA, 1990)

Developer(s): National Education Association

Statements related to inappropriate testing behaviors:

Requires teachers to recognize unethical, illegal, and inappropriate methods of assessment. Fairness, the rights of all concerned, and professional ethical behavior must undergird all student assessment activities, from the initial planning for and gathering of information to the interpretation, use, and communication of the results. Teachers must be well-versed in their own ethical and legal responsibilities in assessment. In addition, they should also attempt to have the inappropriate assessment practices of others discontinued whenever they are encountered. Teachers should also participate with the wider educational community in defining the limits of appropriate professional behavior in assessment.

Teachers who meet this standard will have the conceptual and application skills that follow. They will know those laws and case decisions which affect their classroom, school district, and state assessment practices. Teachers will be aware that various assessment procedures can be misused or overused resulting in harmful consequences such as embarrassing students, violating a student's right to confidentiality, and inappropriately using students' standardized achievement test scores to measure teaching effectiveness.

REFERENCES

Ad Hoc Committee on Confirming Test Results (2002). Using the National Assessment of Educational Progress to confirm state test results. Washington D.C.: National Assessment Governing Board. Retrieved September 14, 2007 from: http://www.nagb.org/pubs/color_document.pdf

Advanced Psychometrics (1993). *Scrutiny!* [Computer software], St. Paul, MN.

Aiken, L. R. (1991). Detecting, understanding, and controlling for cheating on tests. *Research in Higher Education*, 32(6), 725-736

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington DC: American Psychological Association

Angoff, W.H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, 69(345), 44-49.

Anikeeff, A.M. (1954). Index of collaboration for test administrators. *The Journal of Applied Psychology*, 38, 174-177.

Asimov, N. (2007a). School mired in claims of cheating: former university prep teachers issue scathing report – state clamps down. Retrieved July 8, 2007 from *San Francisco Chronicle* web site: http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2007/07/08/UPREP.TMP&tsp=1

Asimov, N. (2007b). Oakland principal in cheating stink quits. Retrieved July 16, 2007 from *San Francisco Chronicle* web site: http://www.sfgate.com/cgi-bin/article.cgi?file=/c/a/2007/07/13/MNGIBR00831.DTL&type=printable

Asimov, N. & Wallack, T. (2007). Cheating the test system. Retrieved May 13, 2007 from *San Francisco Chronicle* web site: http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2007/05/13/CHEATERS.TMP

Associated Press (2007a). Teacher on leave amid probe of alleged cheating. Retrieved May 11, 2007 from *ABC News* web site: http://abclocal.go.com/wtvg/story?section=nation_world&id=5296839&ft=print

Associated Press (2005). Changed test answers lead to teacher resigning. Retrieved February 18, 2007 from *The Ithaca Journal* web site: http://www.theithacajournal.com/news/stories/20050218/localnews/2003002.html

Axtman, K. (2005). When tests' cheaters are the teachers: probe of Texas scores on high-stakes tests is the latest case in series of cheating incidents. Retrieved on January 11, 2005 from *Christian Science Monitor* web site: http://www.csmonitor.com/2005/0111/p01s03-ussc.html

Balassone, M. (2007). Teachers stumble, cheat on state tests. Retrieved August 20, 2007 from *The Modesto Bee* web site: http://www.modbee.com/local/story/45979.html

Bay, L. (1995). *Detection of cheating on multiple-choice examinations*. Paper presented at the annual meeting of the American Educational Research Association.

Belleza, F.S., & Belleza, S.F. (1989). Detection of cheating on multiple-choice tests: an update. *Teaching of Psychology*, 22(3), 180-182.

Benton, J. (2006a). Analysis suggests cheating on TAKS. Retrieved May 23, 2006 from *The Dallas Morning News* web site: http://www.dallasnews.com/sharedcontent/dws/dn/education/stories/052306dnmetcheating.125e559b.html

Benton, J. (2006b). Cheating hasn't hurt Wilmer-Hutchins teachers. Retrieved October 1, 2006 from *The Dallas Morning News* web site: http://www.dallasnews.com/sharedcontent/dws/dn/education/stories/100106dnmetwilmercheaters.344f113.html

Benton, J. (2007a). FW charter school in trouble over TAKS cheating. Retrieved June 22, 2007 from *The Dallas Morning News* web site: http://www.dallasnews.com/sharedcontent/dws/dn/latestnews/stories/061507dnmetcheatinglee.3c44589.html

Benton, J. (2007b). TEA: teacher leaked parts of TAKS test. Retrieved July 16, 2007 from *The Dallas Morning News* web site: http://www.dallasnews.com/sharedcontent/dws/news/texassouthwest/stories/071307dntextaks.3938280.html

Benton, J. & Hacker H. (2007a). Analysis shows TAKS cheating rampant. Retrieved June 3, 2007 from *The Dallas Morning News* web site: http://www.dallasnews.com/sharedcontent/dws/news/dmn/stories/060307dnmetcheating.433e87c.html

Benton, J. & Hacker H. (2007b). Estimated number of cheaters might be low. Retrieved June 3, 2007 from *The Dallas Morning News* web site: http://www.dallasnews.com/sharedcontent/dws/dn/education/stories/060307dnmetconservative.3d3c27f.html

Bird, C. (1927). The detection of cheating in objective examinations. *School and Society*, 35, 261-262.

Bird, C. (1929). An improved method of detecting cheating in objective examinations. *Journal of Educational Research*, 29, 341-348.

Blanchard, J. (2005). Do teachers coach during WASL testing? Retrieved May 16, 2005 from *Seattle Post Intelligence* web site: http://seattlepi.nwsource.com/local/224429_wasl16.html

Bosman, J. (2007). New report clears school of cheating. Retrieved June 27, 2007 from *New York Times* web site: http://www.nytimes.com/2007/06/27/nyregion/27schools.html?_r=1&ref=nyregion&oref=slogin

Bowers, W.J. (1964). *Student dishonesty and its control in college*. New York: Columbia University Bureau of Applied Social Research.

Braun, H. & Qian, J. (2007). Mapping 2005 state proficiency standards onto the NAEP scales. U.S. Department of Education, NCES 2007-482. Retrieved September 6, 2007 from: http://nces.ed.gov/nationsreportcard/pdf/studies/2007482.pdf

Brevard School District (2006). News release retrieved June 25, 2007 from *Brevard School District* web site: http://www.brevard.k12.fl.us/This_Week/releases/pdf/6-30-06%20cocoa%20high%20investigation.pdf

Brock, K.C. (2006). Inquiry targets 20 area schools. Retrieved July 28, 2006 from *Dallas Star-Telegram* web site: http://www.dfw.com/mld/dfw/news/state/15148285.htm

Buckley, J. (2007). Mapping state standards to the NAEP scale. Presentation from the National Center for Education Statistics. Retrieved September 14, 2007 from: http://www.ccsso.org/content/PDFs/NAEPMappingBuckley.ppt

Bunn, D., Caudill, S., & Gropper, D. (1992). Crime in the classroom: an economic analysis of undergraduate student cheating behavior. *Research in Economic Education*, Spring, 197-207.

Bureau of Justice Assistance: Center for Program Evaluation. (2007). Glossary. Retrieved September 11, 2007 from *BJA* web site: http://www.ojp.usdoj.gov/BJA/evaluation/glossary/glossary_h.htm

Burns, J. G. (1988). Computers in class. *Teaching Professor*, 2(7), 2.

Bush, G. (2000). Speech delivered at the 91st annual convention of the National Association for the Advancement of Colored People. Baltimore, Maryland. July 10, 2000. Retrieved September 3, 2007 from Washington Post web site: http://www.washingtonpost.com/wp-srv/onpolitics/elections/bushtext071000.htm

Campbell, D. (1975). Assessing the impact of planned social change. *Social Research and Public Policies: The Dartmouth/OECD Conference*, Hanover, NH; Dartmouth College, The Public Affairs Center.

Cannata, B. (2007). Two Stockton teachers accused of cheating. Retrieved August 28, 2007 from *CBS 13* web site: http://cbs13.com/topstories/local_story_236005351.html

Carlson, K. (2005). WRAL investigates teachers accused of changing EOG test scores. Retrieved January 31, 2005 from WRAL web site: http://www.wral.com/news/4148782/detail.html

Caveon (2008). Caveon case study: Caveon Data Forensics™ Helps Mississippi Department of Education Provide Fair Testing Environment. Retrieved April 25, 2008 from Caveon web site: http://www.caveon.com/case_studies/caveon_case_study_MDE.pdf.

Carnoy & Loeb (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24(4), 305-331.

Center for Evaluation and Assessment (2007). *Test Preparation: Considering the Appropriateness of these Activities: A Professional Development Module for Iowa Educators*. Retrieved August 23, 2007 from the *Iowa Department of Education* web site: http://www.iowa.gov/educate/blogcategory/497/920/

Chadwick, D. (2006). The National Assessment of Educational Progress: The nation's report card. NAEP Short Review ICN. Retrieved September 12, 2007 from the Iowa Department of Education web site: http://www.iowa.gov/educate/index.php?option=com_docman&task=doc_download&gid=2204

Chason, W. M., & Maller, S. (1996). *Utility of the Rasch person-fit statistic in detecting answer copying: A comparison with traditional cheating indices*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Cizek, G. (1999). *Cheating on tests: how to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum.

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.

Cizek, G. J. (2003). Educational testing integrity: why educators and students cheat and how to prevent it. *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators*. Retrieved August 8, 2007 from ERIC web site, No. ED 480 061: http://www.eric.ed.gov:80/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED480061

Cizek, G. J. (2005). *A report to the Texas Education Agency: Review and recommendations related to test security*. Contract 1424. Retrieved March 23, 2008 from Texas Education Agency web site: http://www.tea.state.tx.us/student.assessment/admin/texasreport.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum

Conover, W. J. (1999). *Practical Nonparametric Statistics* (3rd ed). New York: John Wiley & Sons, Inc.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice*, 22(3), 5-11.

Crocker, L. (2006). Preparing examinees for test taking: guidelines for test developers and test users. In S. Downing & T. Haladyna (Ed.), *Handbook of Test Development* (pp. 115-128). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Crown, D.F. & Spiller, M.S. (1998). Learning for the literature on collegiate cheating: a review of empirical research. *Journal of Business Ethics*, 17: 683-700.

Cullen, J.B. & Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system. NBER Working Paper No. 12286. Retrieved September 9, 2007 from NBER web site: http://www.nber.org/papers/w12286

Delaware Code. (2001). Amendment to title 14 of the Delaware code relating to the security of the Delware student testing program and data reporting. Retrieved April 12, 2008 from *Delware State Code* web site: http://delcode.delaware.gov/sessionlaws/ga141/chp081.shtml

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.

Driscoll, D. P. (2000). Clarification of the test administrator's test security and ethics agreement. Retrieved April 12, 2008 from *Massachusetts Department of Education* web site: http://www.doe.mass.edu/news/news.asp?id=655

Driscoll, D. P. (2007). Advisory regarding accommodations for students with disabilities. Retrieved April 12, 2008 from *Massachusetts Department of Education* web site: http://www.doe.mass.edu/news/news.asp?id=3314

Dwyer, D. & Hecht, J. (1994). Cheating detection: statistical, legal, and policy implications. Illinois State University

Education Week (2006). Quality Counts At 10: A Decade of Standards-Based Education. Vol. 25, Issue 17.

Einhorn, E. & Melago, C. (2006). Bronx HS in cheat probe. Retrieved November 20, 2006 from *New York Daily News* web site: http://www.nydailynews.com/news/story/473012p-398009c.html

Elliott, S. (2007). Questions raised about materials that boosted test scores. Retrieved February 4, 2007 from *Dayton Daily News* web site: http://daytondaily.printthis.clickability.com/pt/cpt?action=cpt&title=...%2Flocal%2F2007%2F02%2F03%2Fddn020407citydayinside.html&partnerID=528

Elliott, S. (2007b). With proctors in class, City Day's test scores fall. Retrieved August 24, 2007 from *Dayton Daily News* web site: http://www.daytondailynews.com/n/content/oh/story/news/local/2007/08/16/ddn081607cityday.html

Eltman, F. (2005). Dad accused of giving son regents answers (principal helps his boy cheat). Retrieved June 24, 2007 from *Free Republic* web site: http://www.freerepublic.com/focus/f-news/1432174/posts

Embrey, J. (2007). Texas' education commissioner to resign. Retrieved April 14, 2008 from *American Statesman* web site: http://www.statesman.com/news/content/region/legislature/stories/06/21/21neeley.html

Emons, W.H.M., Sijtsma, K., & Meijer, R.R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research*, 39(1), 1-35.

Eve, R.A. & Bromley, D.G. (1981). Scholastic dishonesty among college undergraduates: Parallel tests of two sociological explanations. *Youth and Society*, 13, 3-22.

FairTest.org (2006). FairTest Examiner. Retrieved June 25, 2007 from *FairTest.org* web site: http://www.fairtest.org/examarts/August%202006/cheating806.html

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.

Ferguson, A. (2007). Utah State University methods of psychology. Retrieved September 11, 2007 from USU web site: http://www.usu.edu/psycho101/lectures/chp2methods/methods.htm

Fessenden, F. (2007). Drop in school test scores raises further questions in Yonkers. Retrieved June 24, 2007 from *New York Times* web site: http://www.nytimes.com/2007/06/24/nyregion/nyregionspecial2/24peoplewe.html

Figlio, D. & Getzler, L. (2002). *Accountability, ability, and disability: gaming the system?* Working paper, University of Florida, 2002.

Fisher, N. I. (1983). Graphical methods in nonparametric statistics: A review and annotated bibliography. *International Statistical Review*, 51(1), 25-58.

Florida Department of Education. (1999). Technical assistance paper: Standards for implementation of student assessment programs in Florida school districts. Retrieved April 11, 2008 from Florida Department of Education web site: http://www.fldoe.org/asp/pdf/ethics.pdf

Frary, R.B., Tideman, T.N., & Watts, T.M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, 2, 235-256.

Frederickson, L. (1984). Teaching test-taking skills. *Social Studies Review*, 23(2), 23-28.

Garcia, N. (2007). Testing honesty: has standardized testing turned some teachers into cheaters? Retrieved May 26, 2007 from *Visalia Delta-Times* web site: http://www.visaliatimesdelta.com/apps/pbcs.dll/article?AID=/20070526/NEWS01/705260328

Garza, C. (2005). Students may have been helped. Retrieved April 19, 2005 from *Star-Telegram* web site: http://www.dfw.com/mld/dfw/news/local/11431968.htm

Gay, G.H. (1990). Standardized tests: irregularities in administering of tests affect test results. *Journal of Instructional Psychology*, 17(2), 93-103.

Ghezzi, P. (2005). Test stress and cheating. Retrieved May 3, 2005 from *Atlanta-Journal Constitution* web site: http://www.ajc.com/metro/content/custom/blogs/education/entries/2005/05/03/test_stress_and_cheating.html

Gonen, Y. (2006). Test-mania fuels cheating at many schools, teachers say. Retrieved November 6, 2006 from *Staten Island Advance* web site: http://www.silive.com/printer/printer.ssf?/base/news/1162819827228590.xml&coll=1&thispage=2

Gonen, Y. (2007). School big in 'test tamper.' Retrieved December 18, 2007 from *New York Post* web site: http://www.nypost.com/seven/12132007/news/regionalnews/school_big_in_test_tamper_892195.htm

Green, J & Winters, M. (2003). Testing high stakes tests: can we believe the results of accountability tests? Civic Report 33, Center for Civic Innovation, Manhattan Institute, Manhattan, NY

Hacker, H. (2005). Cause of diving TAKS scores unclear. Retrieved April 26, 2005 from *The Dallas Morning News* web site: http://www.dallasnews.com/sharedcontent/dws/dn/education/stories/042605dnmetcheating.491c8d28.html

Haertel, E., Thrash, W., & Wiley, D. (1978). Metric-free distributional comparisons. Report. ML-Group for Policy Studies in Education, Chicago, IL.

Haladyna, T.M., Haas, N.S., & Nolen, S.B. (1990). *Test score pollution*. Paper presented at the meeting of the American Educational Research Association, Boston, MA.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher,* 20(5), 2-7.

Haladyna, T.M. & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice,* 23(1), 17-27.

Hall, D., & Kennedy, S. (2006). Primary Progress, Secondary Challenge. Report. Retrieved on September 12, 2007 from http://www2.edtrust.org/NR/rdonlyres/15B22876-20C8-47B8-9AF4-FAB148A225AC/0/PPSCreport.pdf

Hall, J.L. & Kleine, P.F. (1992). Educators' perceptions of NRT misuse. *Educational Measurement: Issues and Practice*, 11(2), 18-22.

Hamilton, L.S. & Stecher, B.M. (2006). Measuring instructional responses to standards-based accountability. Retrieved August 14, 2007 from RAND web site: https://rand.org/pubs/working_papers/2006/RAND_WR373.pdf

Hanson, B.A. & Brennan, R.L. (1987). A comparison of several statistical methods for examining allegations of copying. *ACT Research Report Series No. 87-15*, Iowa City, IA: American College Testing

Hanushek, E.A. & Raymond, M.E. (2004a). Does school accountability lead to improved student performance? NBER Working Paper No. W10591. Retrieved September 1, 2007 from Stanford University web site: http://edpro.stanford.edu/Hanushek/admin/pages/files/uploads/accountability.jpam.journal.pdf

Hanushek, E.A. & Raymond, M.E. (2004b). The effect of school accountability systems on the level and distribution of student achievement. *Journal of the European Economic Association*, 2(2-3), 406-415.

Hanushek, E.A. & Raymond, M.E. (2006). School accountability and school performance. Retrieved September 1, 2007 from Stanford University web site: http://edpro.stanford.edu/hanushek/admin/pages/files/uploads/HanushekRaymond.pdf

Harcourt Assessment (2006). Legal policies. Retrieved September 30, 2007 from *Harcourt Assessment* web site: http://harcourtassessment.com/hai/Templates/Generic/NoBoxTemplat…fTermsandConditionsofSale%2ehtm&NRCACHEHINT=NoModifyGuest#tcpu16

Harcourt Assessment (2007). Terms and conditions of sale. Retrieved September 30, 2007 from *Harcourt Assessment* web site: http://harcourtassessment.com/haiweb/Cultures/en-US/Harcourt/General/LegalPolicies.htm

Harrington-Lueker, D. (2000). When educators cheat. *The School Administrator*, December 2000.

Hatch, J.A., & Freeman, E.B. (1988). Who's pushing whom? Stress and kindergarten. Phi Delta Kappan, 69, 145–147.

Hawaii Department of Education (2007). Fall 2007 Hawaii state science and writing assessments: test security and administration procedures. Retrieved February 24, 2008 from *Hawaii Department of Education* web site: http://downloads.k12.hi.us/assessment/hawaiistateassessment/fall07/TestSecurityAndAdminProcedures.pdf

Hedges, L.V., & Olkin, I. (1985). Statistical Methods for Meta-Analysis. Academic Press Orlando, San Diego.

Hildebrand, J. (2007a). Complaints of test fraud rise. Retrieved June 25, 2007 from *Newsday* website: http://www.newsday.com/search/ny-litest245268180jun24,0,4086000.story

Hildebrand, J. (2007b). Details emerge in Uniondale test scandal. Retrieved June 25, 2007 from *Newsday* website: http://www.newsday.com/search/ny-litest0621,0,2578527.story

Hildebrand, J. (2007c). Security worries ahead of regents. Retrieved June 25, 2007 from *Newsday* web site: http://www.newsday.com/search/ny-liaudi065244073jun06,0,1220786.story

Hill, R. (1998). Using NAEP to compare states' data – while it's still possible. Paper presented at the 1998 Annual Meeting of the National Council on Measurement in Education, San Diego.

Ho, A.D. (2005). *Comparing Score Trends on High-Stakes and Low-Stakes Tests Using Metric-Free Statistics and Multidimensional Item Response Models.* Unpublished Doctoral Dissertation. Stanford University.

Ho, A.D. (2007). Discrepancies between score trends from NAEP and state tests: A scale-invariant perspective. *Educational Measurement: Issues and Practice*, 26(4), 11-20.

Ho, A.D. & Haertel, E.H. (2006a). Metric-free measures of test score trends and gaps with policy-relevant examples. CSE Technical Report #665. University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST), Los Angeles, CA.

Ho, A.D. & Haertel, E.H. (2006b). *(Over)-Interpreting Mappings of State Performance Standards onto the NAEP Scale.* Retrieved September 6, 2007 from: http://www.ccsso.org/content/PDFs/Ho%20Haertel%20CCSSO%20Brief1%20Final.pdf

Ho, A.D. & Haertel, E.H. (2007). *Apples to apples? The underlying assumptions of state-NAEP comparisons.* CCSSO Policy Brief. Retrieved February 23, 2008 from: http://www.ccsso.org/content/PDFs/Ho%20Haertel%20CCSSO%20Brief2%20Final.pdf

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

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(1): 3-17.

Holmgren, E. B. (1995). The p-p plot as a method for comparing treatment effects. *Journal of the American Statistical Association*, 90(429), 360-365.

Horne, L. V., & Gary, M. K. (1981, April). *What the test score really reflects: Observations of teacher behavior during standardized achievement test administration.* Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA. (ERIC Document Reproduction Service No. ED 202 9 18)

Hupp, S. (2005). Teachers who enable cheating to be fired. Retrieved September 19, 2005 from *The Courier-Journal* web site: http://www.courier-journal.com/apps/pbcs.dll/article?AID=/20050919/NEWS02/509190353

Hupp, S. (2006). Teachers gave out ISTEP answers. Retrieved October 11, 2006 from *The Indianapolis Star* web site: http://www.indystar.com/apps/pbcs.dll/article?AID=2006610110503

Illinois State Board of Education (2006). Professional testing practices for educators: Spring 2006. Retrieved January 15, 2008 from *Illinois State Board of Education* web site: http://www.isbe.state.il.us/assessment/pdfs/Prof_Test_Prac_2006_PPT.pdf

Impara, J.C & Foster, D. (2006). Item and test development strategies to minimize test fraud. In S. Downing & T. Haladyna (Ed.), *Handbook of Test Development* (pp. 91-114). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Impara, J.C., Kingsbury, G., Maynes, D., & Fitzgerald, C. (2005). Detecting cheating in computer adaptive tests using data forensics. Paper presented at the 2005 Annual Meeting of the National Council on Measurement in Education and the National Association of Test Directors, Montreal, Canada.

Improving America's Schools Act of 1994, Pub. L. No. 103-382. (1994). Retrieved September 3, 2007 from The U.S. Department of Education web site: http://www.ed.gov/legislation/ESEA/toc.html

Indy Channel (2006). Teacher suspended over ISTEP cheating allegations. Retrieved September 25, 2006 from *The Indy Channel.com* web site: http://www.theindychannel.com/news/9932044/detail.html

Iowa Board of Educational Examiners (2004). *Code of Professional Conduct and Ethics*. Chapter 25 of the Licensure Rules (Iowa Administrative Code). Retrieved August 8, 2005 from the Iowa Board of Educational Examiners web site: http://www.legis.state.ia.us/Rules/Current/iac/282iac/28225/28225.pdf

Iowa Department of Education (2007). Comparing NAEP and ITBS results. Retrieved September 18, 2007 from the Iowa Department of Education web site: http://www.iowaccess.org/educate/ecese/nclb/doc/comparing_naep_itbs.pdf

Iowa Department of Education (2005). Sample board policy. Retrieved August 8, 2005 from the Iowa Department of Education web site: http://www.education.uiowa.edu/itp/documents/DESamplePolicyStatement.pdf

Iowa Testing Programs (2005). *Guidance for developing district policy and rules on test use, test preparation, and test security for the Iowa Tests*. Retrieved August 8, 2005 from the Iowa Testing Programs web site: http://www.education.uiowa.edu/itp/documents/ITPGuidanceDocument.pdf

Iwamoto, C.K., Nungester, R.J., & Luecht, R.M. (1996). *Power of similarity methods and person-fit analysis to detect copying behavior*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Jacob, B.A. (2007). Test-Based Accountability and Student Achievement: An Investigation of Differential Performance on NAEP and State Assessments. Working paper 12817, National Bureau of Economic Research. Retrieved September 1, 2007 from NBER web site: http://www.nber.org/papers/w12817

Jacob, B. & Levitt, S. (2003). To catch a cheat. Retrieved August 8, 2007 from *Education Next* web site: http://pricetheory.uchicago.edu/levitt/Papers/JackobLevittToCatchACheat2004.pdf

Jacob, B. & Levitt, S. (2004). Rotten apples: an investigation of the prevalence and predictors of teacher cheating. Retrieved August 8, 2007 from *Education Next* web site: http://pricetheory.uchicago.edu/levitt/Papers/JacobLevitt2003.pdf

Jan, T. (2006). More teachers accused of cheating. Retrieved August 20, 2006 from *The Boston Globe*, web site: http://www.boston.com/news/local/articles/2006/08/20/more_teachers_accused_of_cheating?mode=PF

Jan, T. (2007). Cheating on MCAS doubles. Retrieved November 5, 2007 from *Boston.com* web site: http://www.boston.com/news/local/articles/2007/11/01/increased_cheating_reported_on_mcas/

Jenkins, L. (2005). With such high stakes, cheating is no surprise. Retrieved August 29, 2005 from the *Union-Tribune* web site: http://www.signonsandiego.com/news/northcounty/jenkins/20050829-9999-1m29jenkins.html

Johnson, T.W. (2006). Elementary school teachers lose certificates. Retrieved July 4, 2006 from the *Baltimore Examiner* web site: http://www.examiner.com/a-167297~Elementary_school_teachers_lose_certificates.html

Johnson, Z. K. (2007). Teacher 'help' crosses the line. Retrieved August 27, 2007 from the *Recordnet.com* web site: http://www.recordnet.com/apps/pbcs.dll/article?AID=/20070823/A_NEWS/708230324

Joint Committee on Testing Practices (2004). *Code of Fair Testing Practices in Education (revised)*. Retrieved August 29, 2007 from APA web site: http://www.apa.org/science/fairtestcode.html

Josephson Institute (2006). 2006 Josephson Institute report card on the ethics of American youth: Part one – integrity summary data. Retrieved September 18, 2007 from: http://www.josephsoninstitute.org/reportcard/

Julian, L. (2007). Performance-pay anxiety: teachers unions may actually do a kind deed for Florida schools. Retrieved December 18, 2007 from the *Orlando Sentinel* web site: http://www.orlandosentinel.com/news/opinion/views/orl-liam1607dec16,0,3096449.story

Kantrowitz, B. & McGinn, D. (2000). When teachers are cheaters. *Newsweek*, June 19, 2000.

Karabatsos, G. (2003) Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education 16:4, 277-298.*

Kentucky Department of Education (2006). Administration code training. Retrieved April 10, 2008 from Kentucky Department of Education web site: http://www.education.ky.gov/KDE/Administrative+Resources/Testing+and+Reporting+/District+Support/Administration+Code+Training+Materials/

Kentucky Department of Education (2007). Mr. Reddy knows all about CATS. Retrieved April 10, 2008 from Kentucky Department of Education web site: http://www.education.ky.gov/SISI_Toolkit/Standard%202/PowerPoints%5CMr.ReddyKnowsAllAboutCATS.ppt

Kher-Durlabhji, N. & Lacina-Gifford, L.J. (1992). Quest for success: preservice teachers' views of "high stakes" tests. Retrieved August 11, 2007 from ERIC web site, ERIC # ED353338: http://www.eric.ed.gov:80/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED353338

Kimmel, E. (1997). *Unintended consequences or testing the integrity of teachers and students*. Paper presented at the annual Assessment Conference of the Council of Chief State School Officers, Colorado Springs, CO, June 1997.

King, L. (2007). Data suggest states satisfy No Child Left Behind law by expecting less of students. Retrieved June 22, 2007 from *USA Today* web site: http://www.usatoday.com/news/education/2007-06-06-schools-main_N.htm

Klein, S. Hamilton, L., McCaffrey, D., & Stecher, B. (2000). What do test scores in Texas tell us? Report, The RAND Corporation, Santa Monica, CA. Retrieved September 5, 2007 from: http://www.rand.org/pubs/issue_papers/IP202/index2.html

Kolen, M.J. & Brennan, R.L. (2004). Test Equating, Scaling, and Linking: Methods and Practices (2nd ed.). New York: Springer-Verlag.

Koretz, D. (1999). Limitations in the use of achievement tests as measures of educators' productivity. Presentation at the *Devising Incentives to Promote Human Capitol National Academy of the Sciences Conference*. Irvine, CA, December 18, 1999.

Koretz, D. (2001). State comparisons using NAEP: large costs, disappointing benefits. *Educational Researcher,* 20(3), 19-21.

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. In J. L. Herman & E. H. Haertel (Eds.), Uses and misuses of data for educational accountability and improvement (104th Yearbook of the National Society of Education, Part 2, pp. 99-118). Malden, MA: Blackwell.

Koretz, D., & Barron, S. (1998). The validity of gains on the Kentucky Instructional Results Information System (KIRIS). Report, The RAND Corporation, Santa Monica, CA.

Koretz, D., & Hamilton, L. (2006). Testing for accountability in K-12. In R.L. Brennan (Ed.), Educational Measurement (4th ed., pp. 531-578). American Council on Education and Praeger Publishers. Westport, Connecticut.

Koretz, D., McCaffrey, D., & Hamilton, L. (2001). Toward a framework for validating gains under high-stakes conditions. CSE Technical Report #551, University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST), Los Angeles, CA.

Kummer, F. & Burney, M. (2006a). How a test-rigging case came to light. Retrieved March 22, 2006 from *The Philadelphia Inquirer* web site: http://www.philly.com/mld/inquirer/14155006.htm

Kummer, F. & Burney, M. (2006b). Test scores drop at Camden High School amid probe. Retrieved June 7, 2006 from *The Philadelphia Inquirer* web site: http://www.philly.com/mld/inquirer/living/education/14758236.htm

Kummer, F. & Burney, M. (2006c). 2 schools' scores plummet. Retrieved June 16, 2006 from *The Philadelphia Inquirer* web site: http://www.philly.com/mld/philly/entertainment/family_guide/14831499.htm

Kvam, P. H. (1996). Using exam scores to estimate the prevalence of classroom cheating. *The American Statistician*, 50(3), 238-242.

Laffont, J.J. & Martimort, D. (2001). The theory of incentives: The principal-agent model. Princeton: Princeton University Press.

Lai, E. & Waltman, K. (2007). High stakes testing and test preparation: examining teacher beliefs and practices. *Center for Evaluation and Assessment*, University of Iowa. Paper presented at the annual meeting of the Iowa Educational Research and Evaluation Association, December 2006. Retrieved June 21, 2007 from the *Center for Evaluation and Assessment* web site: http://www.education.uiowa.edu/cea/documents/Test_Prep_AERA_2007_and_IEREA_2006.pdf

Lee, J. (2006) Tracking Achievement Gaps and Assessing the Impact of NCLB on the Gaps: An In-Depth Look into National and State Reading and Math Outcome Trends. Civil Rights Project. Harvard University, Cambridge, MA.

Levesque, S. (2006). Big Spring's TAKS tests flagged. Retrieved June 9, 2006 from *Abilene Reporter-News* web site: http://www.reporter-news.com/abil/nw_ed_elem_secondary/article/0,1874,ABIL_7951_4762000,00.html

Levine, M.V. & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika,* 53, 161-176.

Levine, M.V. & Rubin, D.B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics,* 4, 269-290.

Ligon, G.D. (2000). Trouble with a capital t. *School Administrator,* 57(11), 40-44.

Ligon, G.D. & Jones, P. (1982). *Preparing students for standardized testing: one district's perspective*. Paper presented at the meeting of the American Educational Research Association, New York.

Linn, R.L. (2005). Adjusting for differences in tests. Paper presented at a Symposium on the Use of School-Level Data for Evaluating Federal Education Programs, Washington DC: The Board on Testing and Assessment, The National Academies, December 9, 2005. Retrieved September 15, 2007 from http://www7.nationalacademies.org/bota/School-Level%20Data_Robert%20Linn-Paper.pdf

Linn, R.L. (2000). Assessments and accountability. *Educational Researcher*, 23(9), 4-16.

Linn, R.L., Baker, E.L., & Betebenner, D.W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, 31(4): 431-435.

Linn, R., Graue, M., & Sanders, N. (1990). Comparing state and district results to national norms: The validity of claims that "everyone is above average". Educational Measurement: Issues and Practice, 9(3):5–14.

Livingston, S. A. (2006). Double p-p plots for comparing differences between two groups. *Journal of Educational and Behavioral Statistics*, 31(4), 431-435.

Loeb, S. & Strunk, K. (2005). Accountability and local control: Response to incentives with and without authority over resource generation and allocation. Retrieved September 1, 2007 from Stanford University web site: http://www.stanford.edu/~sloeb/Papers/LoebandStrunkAccountability.pdf

Loomis, S.C. & Bourque, M.L. (Eds.) (2001). *National Assessment of Educational Progress achievement levels 1992-1998 for reading*. Washington, D.C.: U.S. Department of Education, National Assessment Governing Board. Retrieved September 13, 2007 from http://www.nagb.org/pubs/readingbook.pdf

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Louisiana Educational Assessment Program (2003). *Erasure Analysis Procedures*. Retrieved August 28, 2007 from Louisiana Department of Education web site: http://www.doe.state.la.us/lde/uploads/1607.pdf.

Lucadamo, K. (2006). An early read on test Ed Dept. probe. Retrieved January 12, 2006 from *New York Daily News* web site: http://www.nydailynews.com/news/local/story/382127p-324461c.html

Lyon, K. (2005). Ex-teacher blows whistle in cheating scandal. Retrieved December 23, 2005 from *WCBS-TV New York* web site: http://wcbstv.com/topstories/local_story_357172153.html

Magnuson, P. (2000). High-stakes cheating: will the focus on accountability lead to more cheating. *Communicator,* February 2000.

Marcus, D. (2007a). Why some teachers cheat. Retrieved June 25, 2007 from *Newsday* website: http://www.newsday.com/news/local/longisland/ny-liince0624,0,5126924.story?coll=ny-top-headlines

Marcus, D. (2007b). Experts called to sniff out fraud. Retrieved June 25, 2007 from *Newsday* website: http://www.newsday.com/search/ny-lifrau0624,0,2648649.story

Matter, M.K. (1986). Legitimate ways to prepare students for testing: being up front to protect your behind. *National Association of Test Directors 1986 symposia*. Oklahoma City, Oklahoma City Public Schools.

Maynes, D. (2005). Caveon data forensics pilot report: Pilot analysis of the spring 2005 Texas Assessment of Knowledge and Skills test administration. Retrieved April 26, 2008 from *Caveon* web site: http://www.tea.state.tx.us/student.assessment/admin/security/Caveon_report05.pdf.

McCabe, D., Treviño, L. (1993). Honor codes and other contextual influences. *Journal of Higher Education*, 64, 522-538.

McCabe, D., Treviño, L. (2002). Honesty and honor codes. Retrieved September 30, 2007 from *AAUP* web site: http://www.aaup.org/publications/Academe/2002/02JF/02jfmcc.htm.

McCabe, D., Treviño, L. & Butterfield, K. (2001). Cheating in academic institutions: a decade of research. *Ethics and Behavior,* 11(3), 219-232.

McCollum, D. (2007). Allegations of TAKS cheating. Retrieved June 24, 2007 from *KLTV-7* website: http://www.kltv.com/Global/story.asp?S=6690284

McLaughlin, D., Bandeira de Mello, V., Cole, S., Blankenship, C., Hikawa, H., Farr, K., & González, R. (2002). Comparison of National Assessment of Educational Progress and statewide assessment results: Report to Maryland on 1996 and 1998 assessments. Washington, DC: American Institutes for Research. Retrieved September 14, 2007 from: http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/16/8d/94.pdf

McLaughlin, D., Bandeira de Mello, V., Cole, S., Blankenship, C., Hikawa, H., Farr, K., & González, R. (2002). National Longitudinal School-Level State Assessment Database: Analyses of 2000/2001 school year scores. Washington, DC: American Institutes for Research.

Mehrens, W.A. & Kaminski, J. (1989). Methods for improving standardized test scores: fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice*, 8(1), 14-22.

Mehrens, W.A., Phillips, S., & Schram, C. (1993). Survey of test security practices. *Educational Measurement: Issues and Practices*, 12(4), 5-19.

Meijer, R.R. (1996). Person-fit research: an introduction. *Applied Measurement in Education*, 9(1), 3-8

Meijer R.R. (2005) Using patterns of summed scores in paper-and-pencil and CAT to detect misfitting item score patterns. *Law School Admission Council Computerized Testing Report 02-04,* December, 2005.

Meijer R.R., & Sijtsma K. (2001) Methodology review: evaluating person fit. *Applied Psychological Measurement* 25(2), 107-135

Memphis Eyewitness News (2006). Superintendent: Teachers changed student test answers. Retrieved February 7, 2006 from *Memphis Eyewitness News* web site: http://www.myeyewitnessnews.com/news/local/story.aspx?content_id=372B8005-42B9-47E0-8E66-5EECFF4585D6

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, 21, 215-237..

Mezzacappa, D., Langland, C., & Hardy, D. (2005). Principal accused of cheating on tests. Retrieved April 19, 2005 from *The Philadelphia Inquirer* web site: http://www.philly.com/mld/inquirer/news/local/states/pennsylvania/11412829.ht

Michigan Department of Education (2008). MDE-MEAP Announcement and Information LISTSERV. Retrieved April 23, 2008 from *Michigan Department of Education* web site: http://www.michigan.gov/documents/ListServ_info_111226_7.pdf

Michigan Department of Education & Office of Assessment and Accountability (2007). Professional assessment and accountability practices for education. Retrieved April 23, 2008 from *Michigan Department of Education* web site: http://www.michigan.gov/documents/Ethical_Practices_Document_8-22-05-FINAL_134923_7.pdf

Million, J. (2000). When a principal cheats. Retrieved September 14, 2003 from *National Association of Elementary School Principals* web site: http://www.naesp.org/ContentLoad.do?contentId=231

Minnesota Department of Education (2007). *Procedures Manual for the Minnesota Assessments.* Minnesota Department of Education, 2/26/2007.

Mississippi Department of Education (2006). Mississippi statewide assessment system: roles and responsibilities. Retrieved January 17, 2008 from Mississippi Department of Education web site: http://www.mde.k12.ms.us/acad/osa/security/roles.pdf

Moore, J.L. & Waltman, K. (2007). Pressure to increase test scores in reaction to NCLB: An investigation of related factors. Paper presented at the annual meeting for the American Educational Research and Evaluation Association, April 2007.

Moore, W.P. (1994). Appropriate test preparation: can we reach a consensus? *Educational Assessment*, 2(1), 51-68.

Morris, E. (2007). Manatee teacher's future remains uncertain. Retrieved August 17, 2007 from *Herald-Tribune* web site: http://www.heraldtribune.com/article/20070810/NEWS/708100515

Mosquin, P. & Chromy, J. (2004). Federal sample sizes for confirmation of state tests in the No Child Left Behind Act. Washington, D.C.: American Institutes for Research, NAEP Validity Studies Panel. Retrieved September 13, 2007 from http://www.air.org/publications/documents/MosquinChromy_AIR1.pdf

Mrozowski, J. (2007). MEAP leak forces retest for thousands of students. Retrieved October 16, 2007 from *The Detroit News* web site: http://www.detnews.com/apps/pbcs.dll/article?AID=/20071012/SCHOOLS/710120406/1026

MSNBC (2006). 2 million scores ignored under 'No Child' loophole. Retrieved August 21, 2007 from *MSNBC* web site: http://www.msnbc.msn.com/id/12357165/from/RSS/

Muehlhausen, N. (2007). Investigation details teachers cheating on standardized tests. Retrieved November 12, 2007 from *KSTP Channel 5 Eyewitness News* web site: http://kstp.com/article/stories/S253627.shtml?cat=5

Murphy, K. (2007). Teachers to lose jobs over test help. Retrieved May 20, 2007 from *Inside Bay Area*, web site: http://www.insidebayarea.com/portlet/article/html/fragments/print_article.jsp?articleId=5942288&siteId=181

Muthen, B.O., Khoo, S.T., & Goff, G.N. (1997). Multidimensional description of subgroup differences in mathematics achievement data from the 1992 National Assessment of Educational Progress. CSE Technical Report #432, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles, CA.

National Association of Test Directors (2004). Cleaning up answer sheets. Retrieved August 11, 2007 from http://www.natd.org/Case_3_Cherry_Creek_part_E.pdf

National Center for Education Statistics (NCES) (2007). Item development process. Retrieved September 12, 2007 from the NCES web site: http://nces.ed.gov/nationsreportcard/contracts/item_dev.asp

National Education Association (1990). *Standards for Teacher Competence in Educational Assessment of Students*. Retrieved June 22, 2007 from *Buros Institute* web site: http://www.unl.edu/buros/bimm/html/article3.html

NBC6 (2005). Miami-Dade principal faces cheating allegations. Retrieved April 12, 2005 from *NBC 6* web site: http://www.nbc6.net/education/4370439/detail.html

Neal, D. & Schanzenbach, D.W. (2007). Left behind by design: Proficiency counts and test-based accountability. NBER Working Paper No. 13293. Retrieved September 9, 2007 from University of Chicago web site: http://home.uchicago.edu/~n9na/web_ver_final.pdf

Nelson, T.D. & Schaefer, N. (1986). Cheating among college students estimated with the randomized-response technique. *College Student Journal*, 20, 321-325.

Newberger, E.H. (2003). Why do students cheat? Retrieved August 12, 2007 from web site: http://www.school-for-champions.com/character/newberger_cheating2.htm

Nichols, S.L. & Berliner, D.C. (2005). The inevitable corruption of indicators and educators through high-stakes testing. Education Policy Studies Laboratory, Arizona State University, March, 2005. Retrieved March 12, 2005 from http://www.greatlakescenter.org/pdf/EPSL-0503-101-EPRU.pdf

Nichols, S.L. & Berliner, D.C. (2007). *Collateral damage: how high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

Niels, G. (1996). Is the honor code a solution to the cheating epidemic? Research report from The Klingenstein Center, Teachers College, Columbia University, NY.

No Child Left Behind Act of 2001, Pub. L. No. 107-110. (2001). Retrieved September 3, 2007 from The U.S. Department of Education web site: http://www.ed.gov/policy/elsec/leg/esea02/index.html

Nolen, S.B., Haladyna, T.M., & Haas, N.S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practice*, 11(2) 9-15.

Ohler, A. (2007). Cheating investigation at West Leyden Elementary. Retrieved July 10, 2007 from *News 10 Now* web site: http://news10now.com/shared/print/default.asp?ArID=111329

Ove, T. (2005). Teacher testifies book indicated she could help pupils with test. Retrieved September 29, 2005 from *Pittsburgh Post-Gazette* web site: http://www.post-gazette.com/pg/05272/579553.stm

Parsavand, S. (2007). Corona, Moreno Valley teachers violate state testing rules. Retrieved September 10, 2007 from *The Press-Enterprise* web site: http://www.pe.com/localnews/inland/stories/PE_News_Local_H_cheat08.3df89de.html

Passow, H.J., Mayhew, M.J., Finelli, C.J., Harding, T.S., & Carpenter, D.D. (2006). Factors influencing engineering students' decisions to cheat by type of assessment. *Research in Higher Education,* 47(6), 643-684.

Patrick, K. & Eichel, L. (2006). Education tests: Who's minding the scores? Retrieved June 25, 2006 from *The Philadelphia Inquirer* web site: http://www.philly.com/mld/inquirer/living/education/14898076.htm

Paulhus, D.L. (1991). BIDR reference manual, Version 6. Vancouver, Canada: University of British Columbia, Department of Psychology. Retrieved September 10, 2007 from: http://www.ori.dhhs.gov/education/products/n_illinois_u/datamanagement/dmglossary.html

Pedulla, J.J., Abrams, L.M., Madaus, G.F., Russell, M.K., Ramos, M.A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers.* Boston, MA: Boston College, National Board on Educational Testing and Public Policy

Pellegrino, J.W., Jones, L.R., & Mitchell, K.J. (Eds.) (1998). *Grading the Nation's Report Card: Evaluating NAEP and transforming the assessment of educational progress*. Washington D.C.: National Academy Press. Retrieved September 14, 2007 from http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED446096

Perlman, C. L. (1985, March). Results of a citywide testing program audit in Chicago. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 263 212), pp. 4-5.

Perlman, C. L. (2003). Practice tests and study guides: Do they help? Are they ethical? What is ethical test preparation practice? *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators*, ERIC Document Reproduction Service No. ED 480 062

Peterson, P. & Hess, F. (2005). Johnny can read… in some states. Retrieved June 22, 2007 from *Hoover Institution* web site: http://www.hoover.org/publications/ednext/3219636.html

Peterson, P. & Hess, F. (2006). Keeping an eye on state standards. Retrieved June 22, 2007 from *Hoover Institution* web site: http://www.hoover.org/publications/ednext/3211601.html

Popham, W.J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice*, 10(4), 12-15.

Post, G.V. (1994). A quantal choice model for the detection of copying on multiple choice examinations. *Decision Sciences*, 25(1), 123-142.

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.

Ramsey, J.O. (1991). Kernal smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56(4), 611-630.

Ravitch, D. (2005). Every state left behind. *The New York Times*, November 7, 2005. Retrieved September 13, 2007 from The Brookings Institution web site: http://www.brookings.edu/views/op-ed/ravitch/20051107.htm

Reback, R. (2007). Teaching to the rating: School accountability and the distribution of student achievement. NBER Working Paper No. WP0602. Retrieved September 1, 2007 from NBER web site: http://www.nber.org/papers/wp0602

Reise, S.P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research*, 35(4), 543-568.

Rhodes, S. (2007). WMC-TV Memphis. Germanshire parents sound off over cheating allegations. Retrieved May 22, 2007 from *WMC-TV* web site: http://www.wmcstations.com/global/story.asp?s=6527799&ClientType=Printable

Richards, J.S. (2006a). Cheating is up – among teachers. Retrieved October 22, 2006 from *The Columbus Dispatch* web site: http://www.columbusdispatch.com/news-story.php?story=dispatch/2006/10/22/20061022-A1-01.html

Richards, J.S. (2006b). More districts looking for test violations. Retrieved April 11, 2006 from *The Columbus Dispatch* web site: http://www.columbusdispatch.com/news-story.php?story=dispatch/2006/04/11/20060411-A1-04.html

Riverside Publishing (2006). Sales center: Test security policy. Retrieved September 30, 2007 from Riverside Publishing web site: http://www.riversidepublishing.com/sales/security.html

Roberts, D.M. (1987). Limitations of the score-difference method in detecting cheating in recognition test situations. *Journal of Educational Measurement*, 24(1), 77-81.

Rotherham, A. (1999). Toward performance-based federal education funding: Reauthorization of the Elementary and Secondary Education Act. Progressive Policy Institute Policy Report, April 1, 1999. Retrieved September 3, 2007 from the Democratic Leadership Council web site: http://www.ndol.org/documents/ESEA.pdf

Saupe, J.L. (1960). An empirical model for the corroboration of suspected cheating on multiple-choice tests. *Educational and Psychological Measurement*, 20, 475-489.

Scheers, N.J. & Dayton, C.M. (1987). Improved estimation of academics cheating behavior using the randomized response technique. *Research in Higher Education*, 26, 61-69.

Schmeiser, C.B., Geisinger, K.F., Johnson-Lewis, S., Roeber, E.D, & Schafer, W. (1995). *Code of Professional Responsibilities in Educational Measurement.* Retrieved August 11, 2007 from the NATD web site: http://www.natd.org/Code_of_Professional_Responsibilities.html

Schmidt, E. (2005). Foothills, AZ teacher resigns over cheating incident. Retrieved May 18, 2005 from *Explorer News* web site: http://www.explorernews.com/articles/2005/05/18/education/education01.txt

Shapiro, T. (2005). 'Cheating' probe halts Wai'anae school tests. Retrieved May 9, 2005 from *Honolulu Advertiser* web site: http://the.honoluluadvertiser.com/article/2005/Apr/09/ln/ln03p.html

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3), 591-611.

Shepard, L.A. (1990). Inflated test score gains: is the problem old norms or teaching to the test? *Educational Measurement: Issues and Practice,* 9(3), 15-22.

Shepard, L. & Dougherty, K. (1991). Effects of high-stakes testing on instruction. Paper presented at the annual meeting of the American Education Research Association, Chicago, IL.

Shah, N. (2007). 'I' grades leave schools in limbo. Retrieved July 16, 2007 from *Miami Herald* web site: http://www.miamiherald.com/467/story/171727.html

Singhal, A., & Johnson, P. (1983). How to halt student dishonesty. *College Student Journal*, 17(l), 13-19.

Smith, M.L. (1991). Meanings of test preparation. *American Educational Research Journal*, 28(3), 521-542.

Smith Richards, J. & Riepenhoff, J. (2007). Schools often keep state in the dark. Retrieved October 16, 2007 from *The Columbus Dispatch*, web site: http://www.columbusdispatch.com/live/content/local_news/stories/2007/10/15/FALTER_new.ART_ART_10-15-07_A9_9A85KGB.html?sid=101

Solochek, J. (2007). Teacher accused of FCAT cheating. Retrieved May 2, 2007 from *St. Petersburg Times*, web site: http://www.sptimes.com/2007/05/02/Pasco/Teacher_accused_of_FC.shtml

Sorensen, D. (2006). 2006 state education test security survey results. Retrieved December 17, 2006 from *Caveon* web site: http://www.caveon.com/pr/2006_ed_test_security.pdf

Sotaridona, L. & Meijer, R. (2001) *Two new statistics to detect answer copying*. Research Report RR-01-07, University of Twente Research, Enschede, Netherlands

Sotaridona, L. & Meijer, R. (2002) Statistical properties of the K-index for detecting answer copying. *Journal of Educational Measurement*, 39, 115-132.

Snell, L. (2005). How schools cheat: from underreporting violence to inflating graduation rates to fudging test scores, educators are lying to the American public. Retrieved June 30, 2005 from *Reason* web site: http://www.reason.com/0506/fe.ls.how.shtml

Spencer, B. (1983). On interpreting test scores as social indicators: Statistical considerations. *Journal of Educational Measurement*, 20: 317-333.

SRS1 Software (2007). *SRS1 Cubic Spline for Excel v1.01*. Retrieved December 15, 2007 from the *SRS1 Software* web site: http://www.srs1software.com/

Stanford Policy Repository (2007). Document definitions. Retrieved November 16, 2007 from the *Stanford Policy Repository* web site: http://www2.slac.stanford.edu/policy/definitions.asp

StataCorp. 2007. Stata Statistical Software: Release 10. College Station, TX: StataCorp LP.

STATESUP (2007). Weekly message from state superindendent Randy Dunn 3-28-05. Retrieved March 15, 2008 from Illinois Department of Education web site: http://www.isbe.state.il.us/board/archivemessages/2005/message3-28-05.htm

Stinson, R. (1998). TAAS cheaters meet national standards. *San Antonio Express-News*, September 17, 1998.

Stoneberg, B.D. (2007a). An explanation for the large differences between state and NAEP "proficiency" scores reported for reading in 2005. Paper presented at the 37th Annual National Conference on Large-Scale Assessment, Nashville, TN.

Stoneberg, B.D. (2007b). The valid use of NAEP achievement level scores to confirm state testing results in the No Child Left Behind Act. Paper presented at the NAEP State Service Center Spring Assessment Workshop, Bethesda, MD.

Sturrock, C. (2006). States distort school test scores, researchers say. Retrieved June 22, 2007 from *San Francisco Gate* web site: http://sfgate.com/cgi-bin/article.cgi?file=/c/a/2006/06/30/MNG28JN9RC1.DTL&type=printable

Tatsuoka, K.K. (1984). Caution indices based on item response theory. *Psychometrika*, 49, 95-110.

Texas Education Agency (2005). Texas has zero tolerance for those who cheat students by cheating on test. Retrieved April 26, 2008 from *Texas Education Agency* web site: http://www.tea.state.tx.us/press/takssecurity05.html

Texas Education Agency (2006a). Commissioner names task force on test security. Retrieved April 20, 2008 from *Texas Education Agency* web site: http://www.tea.state.tx.us/comm/page1.html

Texas Education Agency (2006b). Inspector general hired to oversee test irregularity investigations. Retrieved April 20, 2008 from *Texas Education Agency* web site: http://www.tea.state.tx.us/comm/page1.html

Texas Education Agency (2006c). More than 590 schools cleared of testing irregularity allegations. Retrieved April 20, 2008 from *Texas Education Agency* web site: http://www.tea.state.tx.us/comm/page1.html

Texas Education Agency (2007a). Sanctions recommended against three schools and three educators because of testing improprieties. Retrieved October 16, 2007 from *Texas Education Agency* web site: http://www.tea.state.tx.us/press/07iginvestigations.pdf

Texas Education Agency (2007b). Task force on test integrity recommends security enhancements. Retrieved March 26, 2008 from *Texas Education Agency* web site: http://www.tea.state.tx.us/press/taskforcerecomm07.pdf

Texas Education Agency (2007c). Testing audits closed at 88 schools. Retrieved March 28, 2008 from *Texas Education Agency* web site: http://www.tea.state.tx.us/comm/page1.html

Texas Education Agency (2007d). Task force on test integrity recommends security enhancements. Retrieved April 28, 2008 from *Texas Education Agency* web site: http://www.tea.state.tx.us/comm/page1.html

Texas Education Agency (2007e). Test security enhancements planned. Retrieved April 28, 2008 from *Texas Education Agency* web site: http://www.tea.state.tx.us/comm/page1.html

Texas Education Agency (2007f). Commissioner's response to recommendations from the Task Force on Test Integrity. Retrieved April 28, 2008 from *Texas Education Agency* web site: http://www.tea.state.tx.us/comm/page1.html

The Nation's Report Card (2007a). Mathematics report card. Retrieved November 10, 2007 from The Nation's Report Card web site: http://nationsreportcard.gov/math_2007/data.asp

The Nation's Report Card (2007b). Reading report card. Retrieved November 10, 2007 from The Nation's Report Card web site: http://nationsreportcard.gov/reading_2007/data.asp

Thiessen, B. (2004). Aberrant response patterns and person-fit statistics. Retrieved August 12, 2007 from http://homepage.mac.com/bradthiessen/pubs/aberrant.pdf

Thiessen, B. (2006). Educator cheating: classification, explanation, and detection. Thesis Equivalency Project, University of Iowa, June 2006. Retrieved June 22, 2007 from web site: http://homepage.mac.com/bradthiessen/pubs/cheating.pdf

Thiessen, B. (2007). Defining and disseminating policies to address educator cheating: case study. Comprehensive exam paper, University of Iowa, January 2007. Retrieved August 8, 2007 from web site: http://homepage.mac.com/bradthiessen/pubs/format.pdf

Thissen, D. (2005). Linking assessments base on aggregate reporting: Background and issues. Paper presented at the ETS Conference, Linking and Aligning Scores and Scales: A Conference in Honor of Ledyard R Tucker's Approach to Theory and Practice, Princeton, NJ: Educational Testing Services, June 24.

Thomas B. Fordham Foundation (2005). Gains on State Reading Tests Evaporate on 2005 NAEP. Report. Downloaded in November, 2005 from http://www.edexcellence.net/foundation/about/press_release.cfm?id=19

Toomer-Cook, J. (2006). School tests under scope. Retrieved July 3, 2007 from Deseretnews.com web site: http://deseretnews.com/dn/view/0,1249,650212042,00.html

Toppo, G. (2007). School test scandal claims decorated principal. Retrieved January 11, 2008 from *USA Today* web site: http://www.usatoday.com/news/education/2007-12-21-high-stakes_N.htm

Tresaugue, M. & Viren S. (2006). 2 HSID teachers resign in test-cheating probe. Retrieved July 30, 2006 from *Houston Chronicle* web site: http://www.chron.com/disp/story.mpl/front/4080406.html

Turner, D. (2007). Whistleblower out of a job in apparent cheating scandal. Retrieved June 12, 2007 from *WREG-TV Memphis* web site: http://www.wreg.com/global/story.asp?s=6642648&ClientType=Printable

Turtle, J. (2004). How public schools lie to parents and betray our children. *Public Schools, Public Menace*. Retrieved June 16, 2005 from *Press Method* web site: http://www.pressmethod.com/releasestorage/547.htm

United Press International (2005). Teachers ordered cheating. Retrieved March 24, 2005 from *Washington Times* web site: http://washingtontimes.com/culture/20050324-114418-7654r.htm

U.S. Department of Education Institute of Education Sciences (2007). Mapping 2005 State Proficiency Standards Onto NAEP Scales. Retrieved June 22, 2007 from the *NCES* web site: http://nces.ed.gov/nationsreportcard/pdf/studies/2007482.pdf

van der Linden, W. & Sotaridona, L. (2002). *A statistical test for detecting answer copying on multiple-choice tests*. University of Twente Research Report 02-04. Enschede, Netherlands

Wallace, K. (2007). "No Child Left Behind" state tests vary. Retrieved May 30, 2007 from *CBS News* web site: http://www.cbsnews.com/stories/2007/05/30/notebook/main2867441.shtml

Wan, W. (2007). School put on probation after students accused of cheating on AP tests. Retrieved August 24, 2007 from *Washington Post* web site: http://www.washingtonpost.com/wp-dyn/content/article/2007/08/22/AR2007082202645.html

Washington Office of Superintendent of Public Instruction (2005). Memo: critical updates to assessment security. Retrieved December 17, 2007 from Washington Office of Superintendent of Public Instruction web site: http://www.k12.wa.us/Legisgov/2006Documents/CriticalUpdatesAssessmentSecurity.pdf

Waters, B. (2007). Winona ISD had possible TAKS security breach. Retrieved May 1, 2007 from *Tyler Morning Telegraph*, web site: http://www.tylerpaper.com/apps/pbcs.dll/article?AID=/20070501/NEWS01/705010320/-1/FRONTPAGE

Watters, C. (2005). Principal's test scores probed. Retrieved April 13, 2005 from *Rockford Register Star*, web site: http://www.rrstar.com/apps/pbcs.dll/article?AID=/20050222/NEWS0107/502220321/1004/NEWS

W-CBS TV (2006). New York high school accused of fixing test scores. Retrieved November 2, 2006 from *W-CBS TV* web site: http://cbs11tv.com/education/local_story_306101643.html

Wei, X., Shen, X., Lukoff, B., Ho, A.D., & Haertel, E.H. (2006). Using test content to address trend discrepancies between NAEP and California State Tests. ERIC # ED491544. Retrieved September 6, 2007 from: http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/29/00/1f.pdf

Wesolowsky, G.O. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics,* 27(7), 909-921.

White, K. R., Taylor, C., Carcelli, L., & Eldred, N. (1981). *State refinements to the ESEA Title I evaluation and reporting system: Utah 1979- 80 project.* Logan, UT: Utah State University, Exceptional Child Center. (ERIC Document Reproduction Service No. ED 212 087)

WHO TV (2005). Teacher resigns over standardized testing procedures. Retrieved May 4, 2005 from *WHO-TV* web site: http://www.whotv.com/Global/story.asp?S=3302244

Wilk, M.B. & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, 55(1), 1-17.

Wodtke, K.H., Harper, F., Schommer, M. & Brunelli, P. (1989). How standardized is school testing? An exploratory study of standardized group testing in kindergarten. *Educational Evaluation and Policy Analysis,* 11(3), 223-235.

Wolf, R.M. (1992). What can we learn from state NAEP? *Educational Measurement: Issues and Practice*, 11(4), 12.

Wollack, J.A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, 21(4), 307-320.