

Current Guidance For Integrity In Testing

NATIONAL ASSOCIATION OF TEST
DIRECTORS 2005 SYMPOSIUM

Organized by:
Peter Hendrickson, Ph.D.
Everett (WA) Public Schools

Edited by:
Joseph O'Reilly
Mesa (AZ) Public Schools

This is the twenty-first volume of the published symposia, papers and surveys of the National Association of Test Directors (NATD). This publication serves an essential mission of NATD - to promote discussion and debate on testing matters from both a theoretical and practical perspective. In the spirit of that mission, the views expressed in this volume are those of the authors and not NATD. The paper and discussant comments presented in this volume were presented at the April, 2005 meeting of the National Council on Measurement in Education (NCME) in Montreal, Canada.

The authors, organizer and editor of this volume are:

Peter Hendrickson, Organizer and Moderator
Everett Public Schools
4730 Colby Avenue
Everett, WA 98203
(425) 385-4057
phendrickson@everett.wednet.edu

Gregory J. Cizek
University of North Carolina
School of Education, CB 3500
Peabody Hall Room 112
Chapel Hill, NC 27599-3500
(919) 843-7876
cizek@unc.edu

Karen Banks
2113 NE 153rd Ave
Vancouver, WA 98684
360-891-3333
datadetectives@comcast.net

Jim Impara
Caveon, LLC and
Buros Center for Testing
2300 Camelot Court
Lincoln, NE 68512
402 472-8804
jim.impara@caveon.com

Joe O'Reilly, Discussant and Editor
Mesa Public Schools
63 East Main Street #101
Mesa, AZ 85201
(480) 472-0241
joreilly@mpsaz.org

Table of Contents

Personal and Systemic Influences on Integrity in Testing

Gregory J. Cizek.....1

Detecting Cheating in Computer Adaptive Tests Using Data Forensics

James C. Impara, Gage Kingsbury, Dennis Maynes and Cyndy Fitzgerald.....33

A Conceptual Framework For Judging Ethical Violations And Determining Sanctions

Karen Banks66

Discussant Comments

Joe O'Reilly.....82

Personal and Systemic Influences on Integrity in Testing

Gregory J. Cizek

University of North Carolina - Chapel Hill

It seems increasingly common that threats to the integrity and validity of testing are being witnessed, particularly as the stakes associated with passing or failing a test increase. Numerous recent research publications and incidents reported in the popular media indicate that inappropriate test behavior (i.e., cheating) by test takers is on the

rise (see, e.g., McCabe & Trevino, 1996). A 2004 survey of 24,763 high school students conducted by the Joseph & Edna Josephson Institute of Ethics revealed that 62% of high school students admitted to having cheated on an exam within the past 12 months; 83% admitted copying another student's homework and 35% admitted copying an internet document for a classroom assignment at least once. The good news is that, in contrast to some other surveys, the Josephson Institute data suggest that student cheating may be leveling off. The bad news is that, despite their own admissions, 92% of the students surveyed said they were satisfied with their ethics (Josephson Institute, 2004).

At present, however, research and media reports of student cheating appear to be diminishing. In part, this may be due to the fact that bigger (or at least more salient to adults) cheating scandals such as those involving fraudulent corporate earnings reports, unscrupulous investment advisors and others have grabbed the public's attention (see Callahan, 2004). In part, this might also be because copying answers, using crib notes, cutting-and-pasting internet sources to develop a term paper, and the more creative antics of students are so common, so seemingly harmless or amusing, and so expected. What we don't expect is for the gatekeepers to join in on the action.

Displacing some of the news stories about and research interest in student cheating is a focus on cheating by educators. Articles in *Education Week* have

documented incidents of educator cheating across the U.S. including, for example, reports that:

- 7 science teachers in a California school district photocopied the Stanford Achievement Test, 9th Edition (SAT-9) and taught the content it covered;
- teachers at a Chicago elementary school erased wrong answers on students' test booklets and filled in the correct answers, and who filled in answers for questions students had not attempted; and
- 61 principals and teachers distributed answers and corrected students' work as they took tests used by New York state and the New York City district for accountability purposes. (Hoff, 2000)

Attention by researchers to the problem of educator cheating has also increased.

A 2002 report examined results for Chicago Public Schools on the reading and mathematics sections of the *Iowa Test of Basic Skills* (an accountability measure in that district) for all students in 3rd through 7th for the years 1993-2000--yielding an effective sample of approximately 20,000 students per grade per year. The authors of the report derived an index that would be sensitive to unusually large or unsustained score gains and improbable matches in answer strings. The study found "over 1,000 separate instances of classroom cheating, representing 4-5 percent of the classrooms [in the district]" (Jacob & Levitt, 2002, p. 42). A report from the state of Nevada indicated that incidents of student and teacher cheating on that state's test had increased by over 50 percent from the 2002-2003 to the 2003-04 school year (Hurst, 2004).

Some observers have speculated that the rise in cheating in K-12 contexts is unique to education, and they have attributed the increase in cheating by educators to

accountability pressures and external testing mandates such as the *No Child Left Behind Act* (2001). The analysis by Jacob and Levitt concluded that “teacher cheating appears quite responsive to relatively minor changes in incentives” (2002, p. 42) such as those embodied in accountability systems. In the same article that listed the infractions listed above, the author interviewed some opponents of high-stakes testing who conclude that:

“...state accountability rules have increasingly pressured school administrators to prove that their students are learning, often at levels that exceed previous expectations. The main measure has been state- and district-sponsored tests” which have created an “incentive to cheat” (Hoff, 2000, p. 14)

While it is logical to conclude that external accountability pressures certainly have had some unintended negative consequences, those factors alone cannot be culpable for recent increases. Referring to the unique context of licensure and certification testing, Carson (1999) has suggested that the importance associated with test performance assures that controversies (e.g., legal challenges, cheating) are likely to continue. In the context of K-12 education, the problem clearly predates the purported causal factor: the cheating incidents described above predate the passage of the *No Child Left Behind Act* (2001). Three decades ago, the problem was already of concern to at least some observers:

“Teachers cheat when they administer standardized tests to students. Not all teachers, not even very many of them; but enough to make cheating a major concern to all of us who use test data for decision making.” (Ligon, 1985, p. 1)

The problem of cheating by those who administer tests is treated sensationally in the popular press, and perhaps for good reason. Professional athletes, when confronted with examples of their own personal (mis)behavior and the suggestion that such behavior conveys powerful messages to those who observe the behavior--particularly children--can legitimately claim that they are not paid to be role models, but to perform whatever athletic skills they have for the profits of team owners and enjoyment of spectators. Educators, on the hand, can claim no such exemption. The long tradition of teachers-as-role-models is well entrenched in American education. The job description of "educator" includes the requirement that they join as full partners in the task of inculcating in children the values and norms of good citizens. Although no longer as strictly interpreted as acting *in loco parentis*, the clear expectation remains that educators will explicitly teach the attitudes and behaviors deemed desirable for a respectful and harmonious social order.

That expectation is precisely why opposite behavior on the part of educators is so treated so sensationally. We expect students to write formulas on the back of water bottle labels, to use instant messaging to relay answers during tests, and to make liberal use of pre-digested information for a report on Orwell's *1984* from websites such as www.sparknotes.com. An abundance of information on detecting, preventing, and responding to student cheating in both large-scale and classroom contexts (see, e.g., Cizek, 1999, 2003) exists. The "arms race" cycle of student cheating (new methods of cheating are developed, new methods for detecting and preventing them are

developed) will likely always continue, and to some extent, the public and academic attention devoted to the problem will ebb and flow in a similar, cyclical way. What we don't expect--indeed what we find to be so shocking -- is when the teacher who discovers the student filling in formulas on a water bottle label not only fails to confiscate it, but fills in one of his or her own.

In conclusion, the problems of inappropriate test behavior by both test takers and those who administer tests appear to be on the rise. While there has always been concern about ensuring the integrity of testing from threats posed by the individual actions of students, there is more recent and serious concern about threats to testing integrity posed by the actions of educators themselves. The increase in educator cheating cannot be explained simplistically as the result of recent legislation and the introduction of accountability systems.

TWO PERSPECTIVES ON CHEATING

A full understanding of cheating must address two aspects of the problem. For lack of a better way to capture the dichotomy, it may be helpful to label the two aspects as *individual* and *systemic*. The individual aspect comprises much of the current and historic agenda of those who are concerned about cheating. It examines the attitudes, beliefs, and rationales of test-takers or test-givers who cheat. It illustrates the methods used to cheat on tests, term papers, and other academic work; it captures the

independent acts of teachers who change students' answers on high-stakes tests, and it focuses on detecting and preventing individual breaches of academic integrity.

In contrast to the individual perspective, the systemic aspect comprises an interest in the institutional, cultural, social, professional, and other forces that allow, facilitate, or promote cheating. Because the individual perspective is so pervasive and well-documented, the balance of this paper will focus on the systemic perspective and the problem of cheating by those who are responsible for giving tests or overseeing testing programs.

Examples of the Systemic Perspective

The systemic forces that foster cheating are only recently beginning to be documented. In an anonymous personal email note I received, one teacher described the culture at his school:

"Dr: Cizek: I work at a school where cheating is going on by the Principal, Vice President of the union, and The Coordinator in charge of testing. These educators pressure the classroom teachers to cheat. They even coach the teachers how to cheat. This school has the highest test scores in the whole district. The teachers that are forced to cheat are sick and tired of cheating, and scared, but they are afraid if they don't cheat the Principal will make their lives unbearable.

"One of the teachers told the Principal that she had found another district to work for and he yelled at her for not letting him know that she was planning to leave. I think he was more concerned that she was going to break the cheating team. She was afraid of losing her teaching license and tired of cheating, so she decided to save her career and move on. I hate what they are doing to our children.

"I want this to end. I had plans to report the cheating to the Superintendent, but I was told that she is aware of it. My question is how do you stop cheating in your school when the top people are allowing it?"

In my home state of North Carolina, the newspapers have given prominent coverage to the actions of one local district, Johnston County Public Schools, where it has been reported that the school board's goal of raising SAT scores was operationalized in an unexpected way. According to the newspaper account:

“Even as the percentage of NC high-schoolers taking the SAT has risen steadily over the past decade, Johnston County hasn't followed. This year, Johnston's already low rates of participation took an abrupt drop—8 percentage points to 40 percent...the seventh-lowest in the state.

“School officials aren't sure why the drop was so sudden, but the county's approach to the test differs from state and national trends toward encouraging all students to take the test. Instead, Johnston administrators recommend case by case that students take the test, encouraging only those who plan to attend a four-year college or university.” (Fewer take SAT in Johnston County, 2004)

In the first case, it would be hard to describe this kind of systemic pressure as anything less than a school culture aimed at inappropriate score increases--a “cheating culture” to use the term coined by Callahan (2004). In the second case, there are systemic forces at work, but it may be less clear that the actions of the district personnel constitute cheating.

THREE OBSERVATIONS ABOUT EDUCATOR CHEATING

The preceding scenarios help bring out three observations about inappropriate testing behavior on the part of those who administer or oversee tests. The following sections present perceptions and beliefs that contribute to systemic problems, overcoming which will play a significant part in addressing cheating.

Cheating as an Individual Behavior

First, there is clearly a tendency to view cheating as an individual event. Perhaps this is the product of Western (or, at least, American) society, where concern about and respect for the individual is a dominant social, political, economic, and religious norm. In the preceding scenarios, perhaps among of the first questions that come to mind are: “Who *is* that superintendent?” “Which school counselors are doing that?” “Who made that policy?” and so on. In an individual-focused society, we want to know at whom the finger should be pointed.

The tendency to focus on individuals is perhaps most strikingly illustrated in Table 1. In each column of Table 1 there is a list of names. As a pop quiz, try to identify what is the thread that connects each of the names in Column A and the names in Column B.

Table 1
Individual or Systemic Problem?

Column A	Column B
Jeff Mangold	Stan Conte
Gene Monahan	Kevin Shanahan
Hideki Matsui	Tony Torcato
Joe Torre	Felipe Alou

The answer to the quiz comes right out of the sports page headlines. The last names in each column probably clue the answer: Joe Torre is manager of the New York Yankees baseball team; Felipe Alou holds the same job for the San Francisco Giants. But who are the other people on the list? Those in the left-hand column are all also affiliated with the Yankees in various roles--as a strength and conditioning coach, trainer, and outfielder in descending order. Those in the right-hand column have similar positions within the Giants' organization--trainer, massage therapist, and outfielder, respectively.

But what else do the names have in common? According to the sports headlines, many baseball players have admitted to or are suspected of using illegal performance- or growth-enhancing drugs. Among them are two of baseball's biggest all-stars, Barry Bonds of the San Francisco Giants, and Jason Giambi of the Yankees. You have to wonder: Did the guy taking showers next to Barry Bonds every day *really* not suspect that something was fishy? Did the strength and conditioning coach of the Yankees actually think his workout regimen was *that* effective? Was the guy who gives ball players deep tissue massages after every game and practice totally oblivious to "the cream" (a so-called "designer steroid" that is a mixture of testosterone and epitestosterone) slathered on them and blind to the injection sites for "the clear" (tetrahydrogestrinone, a liquid growth hormone that is injected or administered under the tongue)?

If you read the sports pages, you know that there's a lot of "shame on you" talk about Giambi, Bonds, and others. Without diminishing whatever personal responsibility any player--or person--has for his or her own actions, it does seem that there is a lot of culpability to go around. The kind of cheating that allegedly has been committed by some of the leading actors in baseball could not have happened without the tacit approval of a strong supporting cast--a *system* problem.

Another example of an entire system gone bad (sorry, but it's another sports-related example) can be seen in the recent exposé of the University of Georgia (UGA) basketball program reported in the *Atlanta Journal-Constitution*. Almost certainly everyone has heard the stories of college athletes who get a little help to make it through their classes--classes which themselves don't seem to be terribly challenging and bear names that urban legends are made of, like "Introduction to Basket making" and "Exploring Cinema Foods." We suspected that things like this existed, but the truth turned out to be even more incredible than the legends. As part of an NCAA probe into UGA athletics, the university turned over many documents, including the exact final examinations for courses required of the basketball players. A few items from the final examination in a course called "*Coaching Principles and Strategies of Basketball*" as administered in December 2001 are reproduced in Table 2. Three of the 18 bubble-in, multiple-choice items are shown, as is the single, presumably more challenging constructed-response item. And yes, this examination contributed to final grades in the course. Is it just me, or when you begin reading question number five,

when a item begins, “How many halves are in....”, wouldn’t the answer be “2” regardless of how the stem of the item were completed?

Lest my jest be confused here, I want to be clear that the basketball test example is not really about the coach, the team, any individual player, or even about a single institution. Rather, the example is intended to speak to the values and culture in an entire educational system--values that would tolerate the existence of such egregious behavior, or worse, a culture in which such an educational farce would be considered the norm.

Table 2:

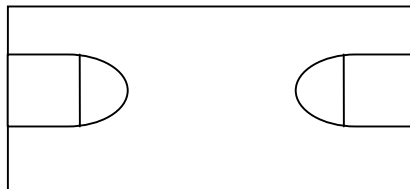
Final Examination in “Coaching Principles and Strategies of Basketball”

- 5. How many halves are in a college basketball game?
 1 2 3 4

- 6. How many quarters are in a high school basketball game?
 1 2 3 4

- 8. How many points does a 3-point field goal account for in a basketball game?
 1 2 3 4

- 16. Diagram the half-court line.



Accountability and High-Stakes Testing

A second (mis)perception is that the influence of accountability systems and the presence of high-stakes tests exert influences that are unique to education. On the one hand, it *is* true that higher stakes and accountability are novel concepts, at least in the K-12 education context. As I have described elsewhere, nascent accountability systems certainly *feel* awkward to the extent that they are new, largely based on student test scores, and have been imposed externally (see Cizek, 2001a). On the other hand, it is easy to see accountability pressures that have long been in place in many (most?) occupations and professions.

On the way to a conference last year, I was seated on an airplane next to a salesman who (if I recall correctly) sold tractors. I was not trying to eavesdrop, but I could hear him mentioning in a cell phone conversation--this was before they closed the aircraft door--that he had to sell X many more tractors this month or he was not going to make his sales goal. I inferred that there were financial stakes associated with not "making the goal."

Perhaps nowhere is accountability more personal and immediate than for persons who wait on tables and rely on gratuities as a major source of income. As a university professor, any increment to my salary each year is based, in large part, on my publications and (thank you, again this opportunity) presentations. Nothing about the nature of providing the noble public service of education is so unique that those who provide that service cannot have its quality or impact judged, or that the differing levels of quality or impact should be unrelated to differing status, compensation, advancement in the profession, and so on.

Related to the accountability-doesn't-fit-all myth is the myth that the severely high-stakes associated with school testing is what make educators cheat. One small-scale study sheds a lot of light on this myth.

The study was an action research project conducted by Sharon Jones, a counselor at Rigby High School in Rigby, Idaho (see Cizek, 2003, for a full account). Apparently the counselor was curious about whether young students would engage in cheating as easily as it appeared that high school students did, So, in 1998 (well before the emphasis

on scientifically-based research) Ms. Jones designed and conducted an experiment. The experiment involved two 2nd grade classes and two 5th grade classes. Each class was given a list of spelling words to study and was told that they would be tested on the words the next day. In one of the 2nd grade classes and one of the 5th grade classes, the students were told that anyone who spelled all the words correctly or made only one mistake would receive a candy bar. For the other two classes (the control group) there was no candy bar offer.

After the spelling tests were administered, Ms. Jones collected the tests and graded them, but she didn't mark grades on the students' papers. Instead, the next day she used a popular method for detecting cheating known as the "I said I didn't grade these papers but unbeknownst to you I really did" technique. She asked the students to grade their own papers, and then she compared the students' self-graded marks with the grades she had recorded previously. Ms. Jones found that, in the "no candy bar incentive" groups, only one student cheated by changing wrong answers to correct ones. In the two experimental classrooms, *all but three of the students cheated*, presumably to get the candy bar. Her conclusion? "Students will cheat if the stakes are high" (quoted in Bushweller, 1999, p. 27).

Ms. Jones' conclusion exemplifies the misapplication of the term "high-stakes" to many current testing situations. It is difficult to imagine much lower stakes than in the spelling quiz situation. Getting a candy bar might be a desirable thing for young children, but the label of "high-stakes" is surely hyperbole. As a reviewer of university

research human participants proposals, it is not likely that the students in this study were deceived into thinking that the reward was any more or any less than the candy bar, and almost certainly no student received a real grade on the spelling quiz. Beyond the misnomer, the study does suggest that even very low stakes--in the right context--can elicit cheating.

In other situations, of course, there *are* somewhat high stakes for students, such as on examinations that are used to help make promotion or graduation decisions. However, even on these “gate keeping” tests, by far the norm is for scores to be evaluated in the context of other information about student performance, for multiple opportunities to attempt the test to exist, and for waiver policies and procedures to be in place. In most state level testing programs where, say, a student “must pass” a state-mandated test in order to be promoted, the vast majority of students who fail the test are promoted nonetheless. For example, one study of a test-based promotion/retention requirement involved a representative sample of Ohio teachers and 5,611 students. Overall, 14% of first time takers of the gate keeping test failed to meet the required standard; the actual retention rate was approximately 2%. Of 801 students who their teachers evaluated as not reading well enough to be academically successful in fifth grade, only 62 (7.7%) were actually retained (Cizek, Hirsch, Trent, & Crandell, 2001).

The term “high stakes” may be equally ill-applied to consequences for educators. The problem of cheating by educators is somewhat remarkable given the rather low-stakes of high-stakes testing. Having searched diligently for one such occurrence, I

have been unable to turn up even a single case of a teacher who was fired because of his or her students' poor performance on high-stakes tests, despite the pervasiveness of rumors that such actions are ubiquitous. The most serious consequences for educators appear to be failure to qualify for salary increments or bonuses, which typify the incentives for performance that many accountability systems have adopted.

Lessening the stakes associated with current accountability systems would not seem to be an effective strategy for reducing cheating. If second graders in rural Rigby, Idaho (969 families, total population = 2,681) will cheat on an essentially no-stakes spelling quiz for a candy bar, then virtually no amount of lessening the stakes associated with tests, grades, and so forth is going to help make noticeable progress in addressing cheating in most education contexts. As Jacob and Levitt (2002) have suggested, the existence of even mild accountability measures seems to provoke misbehavior.

The Definition of Cheating

The third and final observation relevant to the goal of understanding and addressing cheating as a systemic rather than individual problem pertains to the very definition of cheating. There literature in the field of educational measurement is replete with documentation of the inadequacy of assessment knowledge and skill on the part of educators. One author documented what he termed to be a general "apathy concerning testing and grading" (Hills, 1991, p. 540); others have devoted much of their

careers to promoting and enhancing assessment literacy among educators (see, for example, Stiggins, 1991; 1995).

As might be expected in a situation where general knowledge about assessment is generally weak, detailed knowledge about specific assessment concepts is often particularly weak. For example, there is wide disagreement and/or misunderstanding about what is captured by the phrase “teaching to the test.” Much work needs to be done to help define cheating, to promote understanding of *why* certain actions are either unethical, undermine policies, or threaten valid inferences about test performance.

Two years ago now, I received an email note in response to an article I had written on the problem of educator cheating. In the article, I related the true story of how a principal would begin the announcements each morning with a greeting to students via the schools public address system:

“Good morning students and salutations! Do you know what a salutation is? It means ‘greeting,’ like the greeting you see at the beginning of a letter.”

Apparently the students learned the meanings of words like “salutation” from the principal’s daily announcements. They probably never learned that the principal’s choice of words like “salutation” wasn’t random; the words were drawn from the vocabulary section of the state-mandated, norm-referenced test used as a school performance yardstick (Cizek, 2001b, p. 43).

In the article, I didn’t I directly label the principal’s behavior as cheating, although the context in which it was presented and my clear implication was that it was

not right for the principal to disclose the words from the test in advance of its administration and wrong to focus instruction specifically on test items.

A reader of that article--an assistant principal from Ohio--provided comments to me in an email note:

"Mr. Cizek: I just finished reading your article... I totally disagree that a principal teaching vocabulary words as part of daily announcements is a form of cheating. Teaching them what to expect on a test is EXACTLY what we need to do! I would commend the principal for her instructional leadership, not condemn her for cheating!"

As is strikingly evident from this example--and as documented in numerous surveys of educators' perceptions about the propriety of certain behaviors--it is clear that there is not a common and clear conception of what constitutes inappropriate test administration behaviors.

The problem is not all on the part of teachers and administrators, however. Uncertainty regarding what constitutes cheating can be seen in the actions of those who are essentially in the position to define the concept.

As an example, I refer to a recent court case, in which I served as an expert witness for plaintiff test-takers and their medical training school (see Cizek, 2004). The students' scores on an examination had been invalidated by the board representing their profession on the advice of the testing company that provided test development, administration, scoring, and reporting services for the board. At issue was the admitted presence in the college library reserve section, and admitted use by students, of a "study guide" for the board exam. The guide contained approximately 1078 entries

typed up into a 35-page booklet. Each page had two columns for each entry; the first column listed a statement, topic, or question for the entry, the second column listed a key association, related term, or correct answer. Table 3 gives an example of this format.

As the table shows, the examination review material was consistent with the purpose of the examination; that is, to test candidate’s knowledge of anatomy, pharmacology, and basic science facts and principles. An analysis conducted by the testing company found that “approximately 202” of the 1078 entries were similar to items found in the test item pool. Apparently, a working hypothesis that drove the testing company’s (in)validation efforts was that the study guide was compiled as a result of inappropriate access to or so-called “harvesting” of the item pool by prior test takers.

Table 3
Sample Format of Study Guide and Similar Item

Sample Study Guide Entries

5. TCA MOA?	block reuptake of NE, serotonin, and dopamine
18. What increases diffusion?	solubility
49. What happens if you remove lymphatics?	edema
67. Scurvy	Vitamin C deficiency
70. Flexes leg and thigh	sartorius

Item Identified as Matching Study Guide Entry #49

Removal of lower extremity lymph nodes in the treatment of malignancy increases the risk of which of the following:

- A) tumor metastasis
- B) chronic edema
- C) keloid formation
- D) fat embolism

The jury in the case decided in favor of the students. Several facts related to the composition of the study guide were relevant to the decision. Among them were:

- only a small proportion (18.7%) of the entries in the study guide could be identified as “similar” to items in the pool, suggesting that the source of topics for the items was not from harvested tests but from some other source or sources;
- there did not appear to be attempts to use the study guide covertly, as if it were considered to be prohibited test preparation material;
- the testing company did not have well-documented, replicable, explicit procedures for deciding when an entry was “similar” enough to constitute a match with a item in the test pool;
- with the exception of two entries in the study guide, the entries did not contain incorrect options (that is, the entries were not “items” in the sense that testing specialists use the term);
- some of the “answers” suggested in the second column of the document were actually incorrect; and
- the entries that the testing company had identified as “similar” to actual test items were not flagged, highlighted, or in any way distinguishable from the other entries in the study guide.

Were the students cheating by using the guide as a study aid? As is obvious from the testimony I provided in the case, I didn’t think so and still don’t. On the other hand, as a scholar with a special interest in cheating, I am especially concerned about identifying it when it occurs and responding to it appropriately. Invalidation of a test-

taker's score would, to me, seem to be a very mild response in the range of alternatives that are available, particularly in the case of credentialing examinations. Naturally, it concerns me that, despite my conviction in this case, I may have played a part in abetting the cheating. I am sure that all parties in this particular case wondered about whether the actions of the students were best labeled as cheating.

CONCLUSIONS AND RECOMMENDATIONS

It would seem that even the presence of comparatively mild consequences incorporated as part of an accountability system can prompt unethical actions. This should not surprise us, nor should it be cause for alarm. When consequences exist--in any area--there will always be instances of attempts to circumvent the rules to one's advantage. Thankfully, in education, the incidence of such circumvention is low. The overwhelming majority of teachers and administrators apparently work conscientiously and with integrity to help students achieve.

On the other hand, given the explicit expectation that educators should model noble character qualities for students, the education profession has a special obligation to monitor itself for unethical behavior. Preventing and addressing cheating by educators can be accomplished in many ways. Among them, four strategies would seem to be most promising.

First, cheating should be viewed not simply as an individual problem and responded to as such. Rather, the very systems in which the behavior occurs can nurture its persistence or increase, and the cultural norms of the system can sustain and excuse it. A first step in changing the system perspective is by openly addressing the issue. As McCabe and Trevino (1993) have suggested, one reason for the effectiveness of honor codes may have less to do with their explicit identification of proscribed behaviors and associated sanctions, and more to do with the fact that they simply make the issue of integrity a salient one to members of the educational community covered by a code. In short, educators themselves need to relocate test integrity from the responsibility of oversight agencies and testing companies to the business of everyone involved in education.

Second, basic measures to preventing and identifying cheating must be taken. The recent actions of the Houston (TX) Independent School District (HISD) are an important step. The superintendent of that district, in response to a spate of serious allegation of test performance anomalies responded by announcing a number of common-sense steps. Among them, it was announced that HISD will:

- create an Office of Inspector General with oversight and investigation powers related to testing;
- hire several hundred outside monitors to observe testing in progress in schools, both specifically identified schools and randomly-identified campuses;
- initiate a “testing hotline” for reporting inappropriate actions; and
- institute more aggressive reporting of inappropriate behaviors to the State Board of Educator Certification. (HISD, 2005; Manzo, 2005)

Third, in line with more aggressive reporting of inappropriate behaviors must come more serious sanctions. Too often, even when cheating has been clearly identified, the response is feckless. In separate cases in New Jersey, one teacher accused of blatant cheating was allowed to retire; another simply secured employment in another district. It is rare that serious responses to cheating, such as license revocation, are invoked. A recent article documented both the infrequent serious follow-through on cheating, but also the mild consequences. According to the article in *Education Week*, over the nearly five-year interval spanning 1998 through mid-2002, 21 teachers were “caught” cheating on state tests. The inappropriate actions included such things as unapproved prior review of secure tests in advance; tailoring classroom instruction to match specific test questions; directing students during a test to change answers, and so on--actions that would be readily classified as cheating. The report indicated that responsibility for penalties for the infractions was left to individual districts. Further, in some situations the penalty was a “simple admonishment” (Hoff, 2003, p. 27)

Finally, when addressing cheating from a systemic perspective, it is essential to avoid pursuing only those with the least political power. In a personal communication with Donald McCabe, the leading authority on honor codes, we discussed enforcement. Under an honor code, faculty and students commit themselves to both avoid unethical behavior and to report it when they witness it. In college and university settings, even first-offense violations can result in permanent removal from the institution, at least for students. As far as the available evidence is concerned, there does not appear to be any

record of any faculty member ever being sanctioned for an honor code violation, despite the fact that surveys routinely indicate that large percentages of faculty members admit to overlooking cheating – a clear honor code violation. In K-12 education contexts it is surely easier to prosecute an individual teacher for an unethical act than it is to pursue those who bear responsibility for creating or sustaining a culture where that teacher's actions are tacitly condoned.

REFERENCES

- Bushweller, K. (1999). Generation of cheaters. *American School Boards Journal*, 186(4), 24-32.
- Callahan, D. (2004). *The cheating culture: Why more Americans are doing wrong to get ahead*. Orlando, FL: Harcourt.
- Carson, J. D. (2003). Legal issues in standard setting for licensure and certification. In G. J. Cizek (ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 427-444). Mahwah, NJ: Lawrence Erlbaum.
- Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum.
- Cizek, G. J. (2001a). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20(4), 19-27.
- Cizek, G. J. (2001b). Cheating to the test. *Education Matters*, 1(1), 40-47.
- Cizek, G. J. (2003). *Detecting and preventing classroom cheating*. Thousand Oaks, CA: Corwin Press.
- Cizek, G. J. (2004, April). Protecting the integrity of computer-adaptive tests: Results of a legal challenge. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Cizek, G. J., Hirsch, T., Trent, R., & Crandell, J. (2002). A preliminary investigation of pupil proficiency testing and state education reform initiatives. *Educational Assessment*, 7(4), 283-302.

Fewer take SAT in Johnston County. (2004, September 22). *Raleigh News and Observer*, p. B-1.

Hills, J. R. (1991). Apathy concerning grading and testing. *Phi Delta Kappan*, 72, 540-545.

Hoff, D. J. (2000, June 21). As stakes rise, definition of cheating blurs. *Education Week*, 19(41), pp. 1, 14-16.

Hoff, D. J. (2003, November 5). New York teachers caught cheating on state tests. *Education Week*, 23(10), p. 27.

Houston Independent School District. (2005, January 6). HISD to create Office of Inspector General, new controls on testing [HISD press release]. Houston, TX: Author.

Hurst, M. D. (2004, October 6). Nevada report reveals spike in test irregularities. *Education Week*, 24(6), pp. 19, 22.

Jacob, B. A., & Levitt, S. D. (2002, December). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. [NBER Working Paper No. 9413]. Cambridge, MA: National Bureau of Economic Research.

Josephson Institute. (2004). *2004 report card press release and data summary: The ethics of American youth*. Retrieved February 1, 2005 from <http://josephsoninstitute.org/Survey2004/>

Ligon, G. (1985, March), Opportunity knocked out: Reducing cheating by teachers on student tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 263 181).

Manzo, K. K. (2005, January 19). Texas takes aim at tainted testing program. *Education Week*, 24(19), pp. 1, 14.

McCabe, D. L., & Trevino, L. K. (1993). Academic dishonesty: Honor codes and other contextual influences. *Journal of Higher Education*, 64(5), 522-538.

McCabe, D. L. & Trevino, L. K. (1996). What we know about cheating in college: Longitudinal trends and recent developments. *Change*, 28(1), 28-33.

No Child Left Behind Act. (2001). P. L. 107-110, 20 U.S.C. 6301.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan*, 72, 534-539.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, 77, 238-245.

Take the final exam. (2004, March 2). *Atlanta Journal-Constitution*, p. D-2.

Detecting Cheating in Computer Adaptive Tests Using Data Forensics

**James C. Impara, Caveon, LLC and Buros Center for Testing, Gage
Kingsbury, NWEA, Dennis Maynes and Cyndy Fitzgerald, Caveon,
LLC**

Cheating is on the rise in both high school and college settings. In a series of surveys and a review of research on trends in cheating in college, McCabe (2005) of Rutgers University and his colleagues found that in 1961, only 26% of students admitted to copying from another student during a test. That percentage rose to 52% thirty years later in 1991. In a 1999 study, 75% of students admitted to some form of cheating. Cizek (1999) wrote, "...one conclusion from

the trend studies is clear: All agree that the proportion (of cheaters) is high and not going down.” (p. 35)

Over the past 15 years, there has been a strong movement in credentialing testing to move from paper and pencil to computer-based testing (CBT). This trend has been slower to occur in the educational community, especially for high-stakes testing, but there is movement in that direction in both local districts and for state assessment programs. Although providing many advantages for the testing program, students, examinees and users of the test results, this trend has also produced at least one major security problem: an enhanced ability to capture and share test information. Davey and Nering (2002) warn “...at least some of what has been learned over the years about securing conventional high-stakes tests must be updated to meet the new problems posed by CBT administration.” (p. 166). They add, “The danger is not that question pools will be disclosed. As stated, that much is a given – they will be. The danger is they will be disclosed so quickly that economics, logistics and pretest requirements make it impossible...to keep up.” (p. 188) Note that in most educational settings, even when CBT is employed, it is not done in an “on-demand” context. Testing windows are fixed, rather than continuous. This fixed window administration strategy does not necessarily prevent the security risks alluded to by Davey and Nering, but it may help to reduce such risks.

CHEATING

Cheating can occur using a variety of methods such as using inappropriate materials (e.g., PDAs, text messaging, cheat sheets), not stopping when time is called, copying/collusion; by teachers: correcting student errors en masse (erasing wrong answers and inserting correct answers), watching over shoulders and assisting individual students directly, by putting answers on the board, or by obtaining some (or all) test questions in advance of the testing window and using these as “practice” tests. Some of these strategies are made more difficult in a CBT testing mode, but not all. Using computer adaptive tests (CATs) also makes some of these strategies more difficult, but it does not preclude some of these security risks.

Table 1 illustrates how conditions of testing can also influence the type of cheating.

Table 1
Potential cheating methods across different test delivery modes

Cheating/Test Delivery mode	Paper and Pencil	Computer based - linear	Computerized-adaptive
Examinee: text messaging and other forms of two-way communication (e.g., two-way radios)	X	X	X
Examinee: Using unauthorized materials (e.g., calculator)	X	X	X
Examinee: Collusion with another examinee (e.g., copying)	X	X	
Examinee: Proxy testing (having another person take the test)	X	X	X
Examinee: Using brain dumps		X	X

Various approaches can be used to detect cheating. One of the most common approaches is to receive reports of observations by others (someone “rats out” the miscreant); score change anomalies (volatile changes in scores across years – either students, teachers or schools), and data forensics (looking for unusual response patterns, latent response times, erasure analysis, computing collusion indices).

The most common statistical approaches to detecting cheating on CBTs include using latent response times, employing “cheat” programs, and looking at score differences across time (e.g., year-to-year classroom/school/district score changes within a grade or for a student cohort, or item drift). Caveon’s approach employs some of these same strategies, but enhanced in several ways. These strategies have not been used previously with CATs.

PURPOSE

The purpose of the present paper is twofold. The first section of the paper discusses data forensic methods for detecting test fraud using indicators of aberrance (unusual response patterns and unusual patterns in item response times) and the second half provides results of the application of these measures to an educational assessment program that uses computerized adaptive testing (CAT).

NWEA and Caveon conducted a study to investigate the potential and power that data forensics methods, founded in measures of aberrance¹, collusion², and score volatility³ have for detecting exam fraud in an educational CAT environment. The primary study goals were to assess detection rates in live data and to assess the impacts of aberrant test taking on the test results. Secondary study goals were to evaluate the security strength of CAT as a means of test delivery and to determine whether measurement of aberrance in a CAT (which is inherently adapted to the examinee's ability and theoretically a near-optimal test) is possible and what it might mean if it were discovered.

In order to assess detection rates of test fraud in live data, the data were seeded by NWEA staff (and unknown to Caveon) with known instances of anomalous test results⁴. The data forensics analyses were performed by Caveon to present the impact of aberrant and collusive test-taking on the test results.

The results are presented in two stages. In stage one the overall results are described based on a particular, and new, approach to analyzing data to look for data anomalies. The second stage represents how accurate this approach was to

¹ Aberrance is observed when a subject answers the test questions in a manner that is inconsistent with demonstrated knowledge and behavior. Examples are inconsistencies in the amount of time taken to respond to test items, and answer selections that are inconsistent with a student's demonstrated ability on other test items.

² Collusion occurs when examinees share answers for the test items either during or before the test. It also manifests itself when an educator provides the same answers to test items to multiple students. Statistically, collusion indicates that the tests are not being taken independently.

³ Score volatility is measured when a student retakes the test and demonstrates an extreme score change. When the change is so extreme the practitioner disbelieves that the result is due to chance and may believe that the result is due to cheating.

⁴ Note that the seeding (usually changing a wrong response to a correct response) did not include adapting the routing of the test. Thus, responses to subsequent items were not affected as would have been the case had the "cheating" been done by the student. This may have had an impact on the ability of the data forensics to detect the cheating that was seeded by NWEA.

detecting “known” cheating situations. Before discussion of the results, some terms are defined and the nature of the data is described.

ABERRANCE

Aberrance refers to a test result that does not conform to the test response model. Because there are many types of non-conformance this term lacks a precise definition. Some types of non-conformance include wild guessing (as opposed to educated guessing), poor test preparation, mis-keyed test questions, and cheating (or pre-knowledge of some or all of the test content).

Identifying response aberrance begins with an examination of the individual responses in context of all the responses. A single response to a single test question cannot be construed as aberrant or not, except in the sense that the response may be so improbable that it causes the test administrator extreme surprise. This concept of surprise (due to observing extremely improbable events) is an essential aspect of aberrance. Item Response Theory (IRT) models provide the statistical framework for objectively measuring the probability of a set of responses and from probability to “surprise” when the item responses do not conform to the testing model.

Different kinds of non-conformance, or aberrance, are in reality different patterns of unusual or improbable responses. A method that attempts to evaluate

patterns of responses must be capable of differentiating between the different patterns as they relate to the observed responses.

Another aspect of aberrance is that not all responses will necessarily be improbable. We are then left with the situation that each observed response has a different probability and the numbers of improbable responses directly correlate with our notion of surprise or non-conformance. A single improbable response should rarely convince us (or provide sufficient evidence) that the test is being taken inappropriately. An exception might be the case where a person does extremely well answering nearly all questions correctly but then answers an easy question with a very improbable incorrect response (which could also be termed a blunder). Similarly, when pilot testing of new items is done by embedding them in an operational test (as compared to stand-alone pilot testing), test behavior on the pilot test items that is not consistent with performance on the operational items can be indicative of anomalous performance.

The aggregated evidence of conformance versus non-conformance needs to be evaluated in order to convince us that a test was taken inappropriately. It is only by considering all the responses in context and then ordering those responses by the degree of surprise (or improbability of occurrence) that a particular set of responses can be viewed as “aberrant.” In other words, aberrance is a property derived from the individual responses, based on all the responses.

For the current purposes, aberrance is defined as the number (or percent) of improbable responses and the degree of improbability associated with those responses in the observed test responses. It is well known that the sample mean and sample standard deviation are not resistant to outliers and influential observations. In the same manner, outliers or aberrant responses heavily influence the estimation of theta in the IRT model. This influence has the potential to mask the true aberrant responses, making detection of aberrance (or responses that generate “unusual surprise”) difficult.

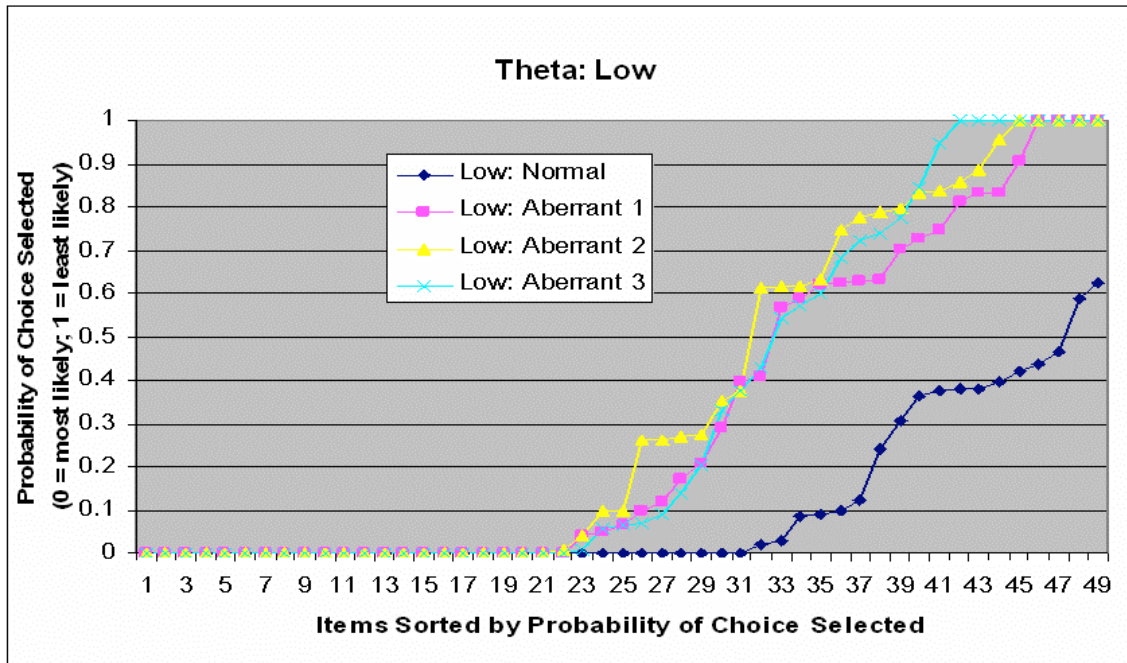
If a test has psychometric integrity, little or no aberrance will be seen in the test responses of the individual who responds to the test fairly and honestly. The cheater can be viewed as a test-taker who has an unfair prior knowledge (or knowledge gained during the exam) of the test content. If the cheater has gained access to the entire test content, then it is unlikely that response aberrance models will detect this behavior. Instead item response latency aberrance models will be required. On the other hand, if the cheater has gained access to less than 100% of the content, then this individual can be viewed as responding differently to the questions, depending on prior knowledge. The individual will respond to the questions with prior knowledge at a higher level of theta than to the questions without prior knowledge. The cheating model is then a model where two levels of theta are presumed to be exhibited (i.e., the response pattern will be bi-modal in terms of estimating theta).

With sufficiently large data sets, even unlikely patterns will show up from time-to-time. An example is the lucky guesser who is able to guess a significant number of correct answers. And another example is the individual who makes a lot of “stupid” mistakes (e.g., who may have accidentally got off line on the response sheet). In both of these cases the actual responses will not reflect the test taker’s actual knowledge. These are such low probability occurrences that they do not merit separate models, but they will be present in large data sets by chance alone.

Caveon believes that aberrance in response patterns and response latencies (when available) are one of the better indications of cheating and item theft. For computer-based testing, six different measures of aberrance are used in the analysis, three that look at the answers examinees select, and three that look at how long it takes the examinee to answer the question. By combining all six into a single aberrance value, we can get useful information on the rates of security problems.

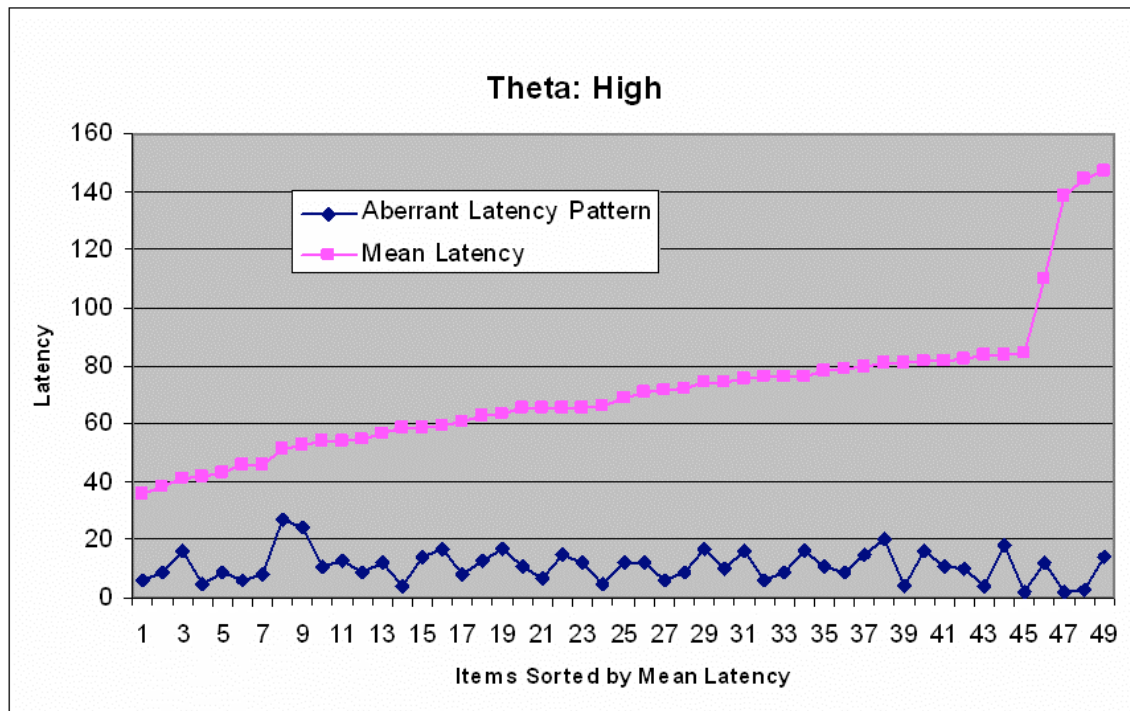
But let’s first show two examples of Caveon’s aberrance measures. These examples come from the certification and licensure arena. Figure 1 shows the response patterns for 4 individuals with the same low score. It indicates three aberrant tests and a normal one. Although the statistics are far from simple, unusual or improbable selections of answers to items make it easy to separate the aberrant tests from the normal ones. One can see such differences at every ability level except the very highest.

Figure 1



With computerized tests we are able to add in item response-time measures to the aberrance analysis. Figure 2 shows a test record for an individual who scored very high on the test even though the amount of time to respond to each question was too brief to allow him or her to answer the questions in a normal way. The amount of time taken was uncorrelated ($\tau = -.04$) with the amount of time the test takers as a whole took to answer the questions. In the figure, latency is the amount of time taken to respond to each question (sorted by average latency for the entire group of test takers). This is the line with boxes. The individual examinee is shown to have a very quick response time, less than 20 seconds for most items, in comparison to the average examinee.

Figure 2



At Caveon we have found aberrance to be a useful tool in tracking differences across the entire testing population. The statistical models used generate a 5% aberrance rate if there is no actual aberrance in the test results. Using this as a baseline comparison, we can evaluate and compare across localities (e.g., regions, districts), and even specific testing locations (e.g., schools, classrooms, testing centers). There are a few terms that need to be defined in order to be able to interpret the data presented below. They are presented in Appendix A.

Table 2 provides an illustration of statewide rates for a sample of 20,661 educational tests. What is most alarming is that a full 17% of the tests show as aberrant in this case. (Detailed data are provided for all tests and for the tests

with the highest aberrance levels.) Some explanation of the data is in order.

Aberrance is defined as above. High scores are passing scores and low scores are failing scores (in this case, the passing score was set arbitrarily at the median because the test didn't have a passing score). The cheating index and the piracy index are both combinations of the six aberrance measures explained above; higher values are indicative of higher likelihoods of either cheating or piracy.

Table 2
District Summary

District	Tests	Mean Percentile	Pass Rate %	Pass Rate Index	Aberrance Rate %	Aberrant High Score Tests	High Score Tests	High Score Aberrance Rate	Cheating Index	Aberrant Low Score Tests	Low Score Tests	Low-Score Aberrance Rate	Piracy Index	Comments
Overall	20661	0.50	49	0.0	17	1753	10221	17	0.0	1815	10440	17	0.0	
107	1614	0.46	45.1	-3.6	19.3	141	728	19	1.3	171	886	19	1.2	
263	1463	0.52	52	1.4	20	149	761	20	1.5	144	702	21	1.9	
444	1386	0.53	55.1	4.9	23.7	184	764	24	7.2	145	622	23	4.5	The elevated high-score and low-score aberrance rates make this anomalous. These rates are indicative of cheating.
912	1028	0.43	38.9	-	13.7	63	400	16	0.1	78	628	12	0.0	
412	820	0.42	37.6	-	17.6	57	308	19	0.6	87	512	17	0.2	
719	807	0.55	56.1	3.9	18.5	81	453	18	0.5	68	354	19	0.7	
122	670	0.51	50.3	0.2	26	79	337	23	3.0	95	333	29	7.6	The elevated high-score and low-score aberrance rates make this anomalous. These rates are indicative of cheating.
988	614	0.50	46.6	-0.8	17.3	44	286	15	0.1	62	328	19	0.6	

We can see similar results when looking at individual schools in Table 3. The school at the top, labeled 4379, has a very high overall aberrance rate of 36%, mostly for tests with high scores. The Caveon Cheating Index of 12.4 indicates the probability of this result occurring by chance as less than .00000001!

**Table 3
School Detail**

School	Tests	Mean Percentile	Pass Rate %	Pass Rate Index	Aberrance Rate %	Aberrant High Score Tests	High Score Tests	Aberrance Rate	Cheating Index	Aberrant Low Score Tests	Low Score Tests	Aberrance Rate	Piracy Index	District	Comments
8500	807	0.55	56	3.9	18	81	453	18	0.5	68	354	19	0.7	719	
9456	789	0.55	59	7.1	14	71	464	15	0.1	43	325	13	0.0	777	This is a high pass rate.
7001	545	0.50	50	0.1	13	35	273	13	0.0	34	272	13	0.0	289	
5514	501	0.38	33	12.9	13	24	166	14	0.1	39	335	12	0.0	912	
8052	442	0.52	49	-0.1	24	50	217	23	2.0	55	225	24	2.6	851	
4379	383	0.56	61	5.4	36	81	234	35	12.4	56	149	38	10.5	444	This pass rate is in the presence of high aberrance for high and low scores. This may indicate test coaching at the school.
8849	372	0.53	52	0.6	11	22	195	11	0.0	18	177	10	0.0	777	
9621	359	0.42	38	-5.0	26	32	136	24	1.6	60	223	27	4.1	528	A high amount of low score aberrance with low pass rates. Most likely the students are unprepared to take the exam.
7047	345	0.53	56	1.8	22	45	193	23	2.0	30	152	20	0.7	444	
9453	324	0.50	50	0.1	18	24	163	15	0.1	35	161	22	1.1	777	
9161	317	0.63	68	11.0	17	30	217	14	0.0	25	100	25	1.7	875	Very high pass rate.
5117	306	0.44	41	-2.5	11	17	126	13	0.1	18	180	10	0.0	940	
6669	295	0.41	35	-6.3	21	27	103	26	2.1	34	192	18	0.3	107	

Using statistical models, it is also possible to discover and verify proxy testing (having someone else take the test instead of the person who registered to take the test), copying and other forms of collusion. As an example, if a proxy testing service is operating (a fairly frequent occurrence in some foreign certification and admissions testing programs), then scores for different examinees should appear too similar and have other suspicious patterns (this is an unlikely scenario in an educational setting because the teachers know who the examinees are!). Similarly, we should be able to identify individuals who are taking tests collectively as a group, with or without the help of an instructor or other third party. Or if a group is using similar crib/cheat sheets, unauthorized Web resources, or some other means of working together other than copying from each other's test papers.

COLLUSION ANALYSES

Table 4 shows a cluster of seven tests with different examinee IDs. These data are from a sample of examinees who took a certification test. Also shown are their scores and the date and time of the test. Caveon's collusion statistic identified the tests as matching too closely to have occurred by chance. And notice the pattern of testing. All tests occurred on the same day. Furthermore, mostly each test started at about 20-30 minute intervals. This is likely a situation where a proxy test taker is operating. As noted above, this type of proxy test

taking is not something that may be of high concern for most educational tests. However, as more and more high stakes tests come on line this may become more of a problem, especially when proctors are not the students’ teachers..

**Table 4
Proxy Testing Illustration**

	A	B	C	D	E	F	G	H	I	J	K
1	Examinee	Test	Site	Country	Date	Time	Score	Prob			
2	283	101	Site12	US	1/15/2003	1:04:18 PM	0.77	246			
3	405	101	Site 4B	US	5/24/2003	3:17:44 PM	0.87	143	Looks like a proxy test taker.		
4	351	101	Site 4B	US	5/24/2003	3:47:03 PM	0.88	258			
5	860	101	Site 4B	US	5/24/2003	4:12:16 PM	0.88	5029			
6	446	101	Site 4B	US	5/24/2003	4:48:52 PM	0.85	88			
7	440	101	Site 4B	US	5/24/2003	5:16:50 PM	0.83	5029			
8	123	101	Site 4B	US	5/24/2003	4:09:46 PM	0.88	85			
9	559	101	Site 4B	US	5/24/2003	4:46:14 PM	0.82	85			
10	756	101	Site 17	US	1/24/2003	2:34:00 PM	0.85	2134			
11	659	101	Site 17	US	4/11/2003	9:42:30 AM	0.85	2134			

For paper-and-pencil tests, the collusion analysis is equally effective, identifying traditional copying or instances where a teacher may be systematically erasing and changing the answers of students in his or her class.

RETAKE ANALYSES

Although not typically a problem in many educational settings, most certification testing programs have policies that permit, but impose conditions on, retakes. It’s important to track violations of these policies and to look at large gains or losses in test scores as tests are retaken. Both the violations of retake policies and retake gains and losses – what we refer to as volatile retakes – may

be indicators of attempts to cheat or steal questions. Table 5 shows a list of volatile retakes from 9 examinees, retakes where the scores have changed, up or down, more than they should (based on client expectations and reliability estimates). Two of these examples are discussed. First, notice that Examinee 8879 at the bottom scored 85% after a score of 0% on the previous exam. Examinee 6343 retook the test after passing with a very high score of 92%. His/her score on the retake was only 5%. The first example may be an examinee who was either trying to familiarize him or her self with the test when they took it the first time, or who was trying to steal items. The second example suggests the examinee was trying to memorize items during the second attempt with no intention of getting a high score (after all, he/she had already passed).

Examining the results of retakes and volatile retakes becomes relevant in the following educational contexts. First, when a test, like a graduation test, is given several times a year, students who retake may be retaking in violation of policy (policy may restrict retakes once the test is passed). Students in this situation may also demonstrate volatile retakes that should be flagged because they may be trying to gain knowledge of the test content to share with their friends who did not do well on their first attempt or they may have gained knowledge from their friends who shared information from their earlier testing experience. Second, when the year-to-year scores within particular classrooms, schools, or districts demonstrate volatile retakes. Large year-to-year changes (in either direction) should raise a flag that suggests a variety of explanations (e.g.,

change in student population, change in administration/teaching emphases, cheating).

Table 5
Examinee Report

	A	B	C	D	E	F	G	H
1	Examinee ID	Test Site	Country	Passed	Score	Previous Score	Difference	Z-score
2	2022	1111	MEX	1	0.92	0.50	3.30	
3	8271	2222	JPN	1	0.90	0.43	3.83	
4	5723	2222	JPN	1	0.88	0.50	3.00	
5	6183	3333	USA	1	0.95	0.57	3.42	
6	6273	3333	USA	0	0.17	0.40	-3.65	
7	5778	3333	USA	0	0.00	0.67	-9.40	
8	6343	6666	AUS	0	0.05	0.92	-7.74	
9	8879	7777	SGP	1	0.85	0.00	4.23	

EFFECTS ON TEST PERFORMANCE

Caveon's data forensics can determine the effects of aberrance, collusion, and volatile retakes on the performance of the exam by also looking at item drift. Item drift is probably a misnomer given how item exposure occurs today. Drift implies a gradual degradation of an item's effectiveness. What is more often observed is not gradual; in some cases, it's immediate. "Rapid Deterioration" may be a more appropriate term. Instead of casually planning item replacement schedules; testing programs may need to detect and replace items (or entire test forms) when the compromise to item integrity is discovered. This can be problematic in a setting with a relatively wide administration window and a

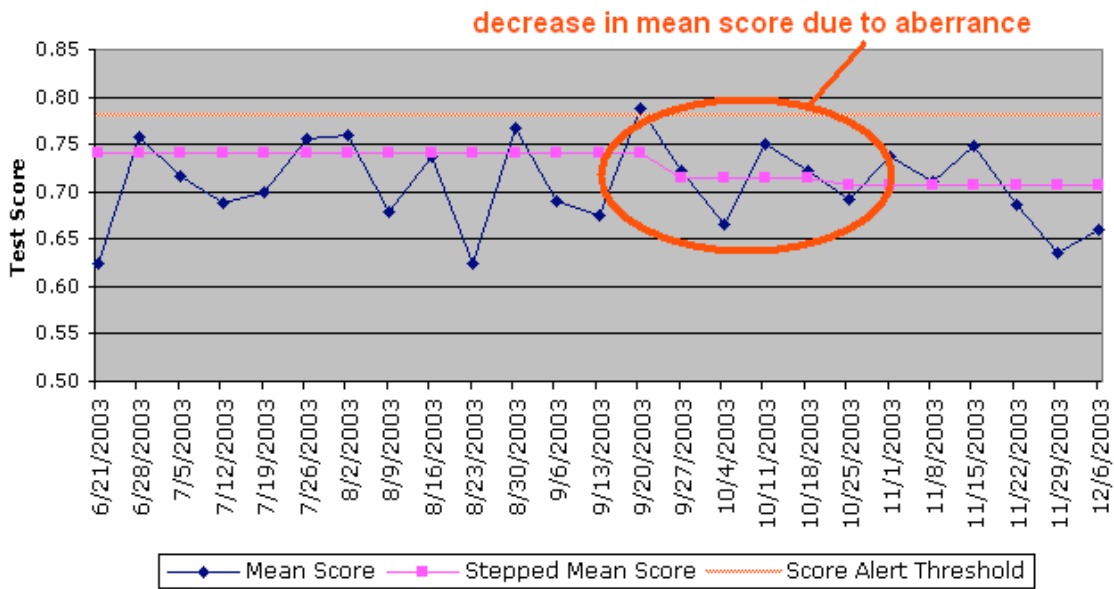
relative small item pool. One state that has recently instituted a computer-based testing program in schools, for example, expects its testing window to be about three weeks and there are fewer than 150 items in the pool. There is substantial risk that virtually all items will have been exposed by the end of the window. This has serious validity implications and it has implications for year-to-year equating and other important aspects of the program.

Caveon has been researching the impact of changes in test performance and aberrance over time. Figure 3 shows an example of a strong relationship between aberrance and test score changes. In this case, an increase in aberrance in a continuously administered certification test over 4 weeks led to a real decrease in test scores during that period. This suggests that the aberrance was indicative of a group of examinees who were likely memorizing test items.

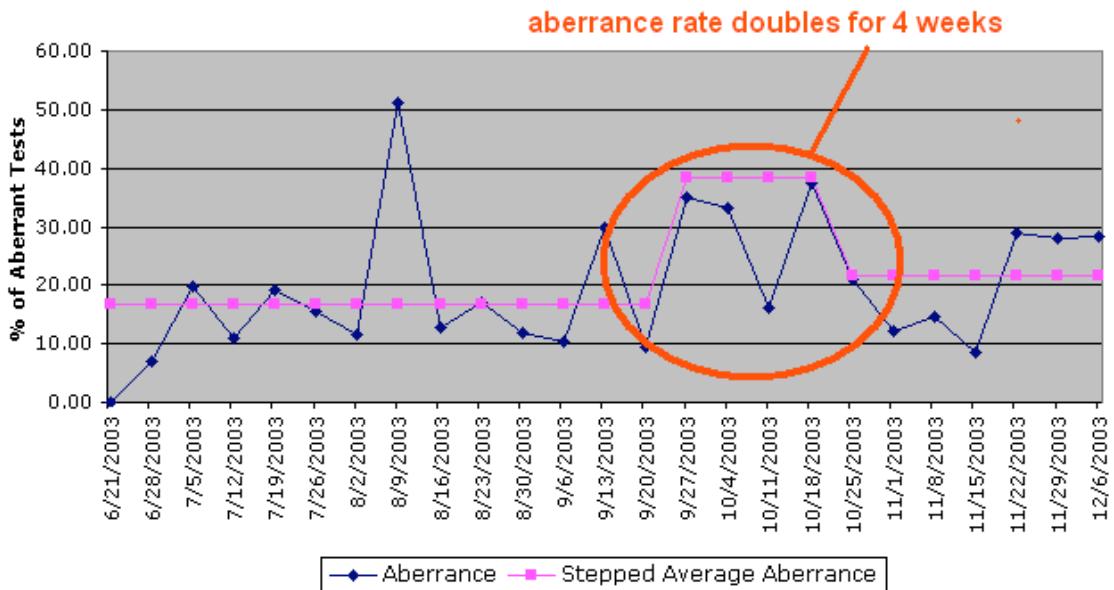
It is a short step from there to also discover the real performance changes in items. Hambleton (2004) presented a method based on observed changes in typical item statistics. These efforts should be viewed as more than looking at item drift and test performance at particular points in time, but rather, using new methods to monitor exam and item performance continuously.

The above is a general overview, using certification and licensure data along with the NWEA educational data, of how various measures of aberrance can be used to detect test fraud. The remainder of the paper provides results of the application of these measures to an ongoing educational assessment program that uses computer adaptive testing.

Figure 3
CVN-101 v1 - Test Score/Time



CVN-101 v1 - Aberrance/Time



Four tests using test results from 20,661 test administrations were used to check for evidence of test security compromise, including cheating and piracy. The reporting period of the analysis was from March 1, 2004 to June 30, 2004.

Selected student records were modified to reflect a very modest amount of cheating behavior.

IDENTIFICATION OF MODIFIED RECORDS

In order to test the approaches used to identify cheating, two school districts, five schools, and 3283 students were identified as “cheaters.” The response records for these individuals were modified to measure the data forensics detection rates.

Ten percent of the items from the CAT item bank were marked as exposed. Whenever a test record from the modified set contained one of the exposed items, that item response was changed so that it was correct (unless it was already correct, then no change was made). As noted in a footnote above, only that item was changed, thus, the conditions did not mimic the actual testing situation in which the examinee might have been routed to a different next item potentially resulting in a more readily detectable aberrant response pattern. The number of items given to the students on the CAT tests was at least 50, which was also the modal test length. Therefore, on average up to 5 test questions would have been modified on each test depending on how many of the five items were answered correctly by the examinee. Review of the CAT data shows that, for most items, the probability of responding correctly to a question is 50%. Therefore, the responses for 2.5 questions on average would have been changed

from incorrect to correct. This is a very small amount of change and cheating in such low incidence situations is very difficult to detect.

The summary of the results of Caveon’s data forensics analysis are:

- One of the two school districts as having a high cheating rate was identified.
- None of the 5 schools as having a high cheating rate was identified.
- Forty one (slightly over one percent) of the 3283 students was identified as being potential cheaters.

In this analysis Caveon’s Data Forensics examined four types of test security risk:

- 1) Collusion -- answer copying, collaboration and communication during testing such as text messaging, teacher coaching and proxy test taking,
- 2) Cheating -- having advanced knowledge of some or all of the exam content (as was intended in the seeding of “cheating” data),
- 3) Piracy -- stealing test items by memorization or technology, and
- 4) Volatile Retakes -- extreme score changes between successive test administrations.

Table 6 lists the tests and some general test details:

Table 6: Overall Test Summary

Test	Security Assessment	Number of exams	Testing sites	Pass rate ⁵
NWEA-394	Moderate/High	5056	16	49
NWEA-396	Low	5304	48	50
NWEA-704	Low	5138	76	49
NWEA-708	Low	5163	26	49

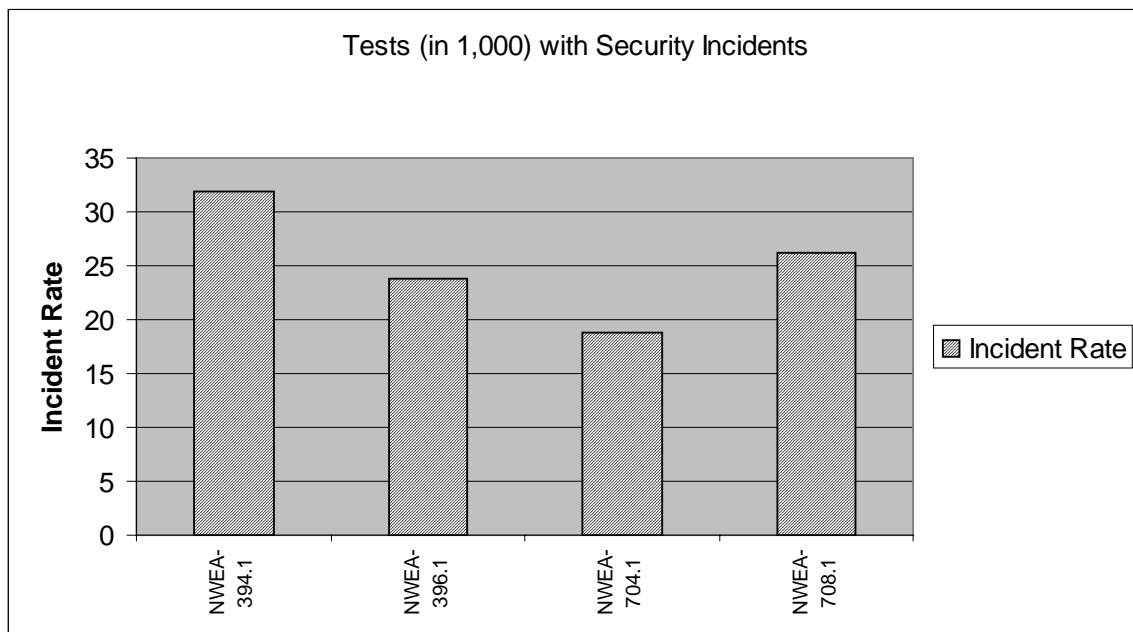
⁵ Pass rates were not defined for this test; to demonstrate the analysis a passing score was set arbitrarily at the median score. Thus, the pass rates should be close to the 50th percentile, subject to the granularity of the score distribution.

One test, NWEA-394, is deemed to have significant risk of test compromise. The reasons for this determination are discussed below.

SECURITY INCIDENT OVERVIEW

Figure 4 compares the security incidence⁶ rates for each of the tests. This figure shows the proportion of tests for which any security incident was identified. A security incident occurs when the scores demonstrate aberrance, collusion, piracy, or volatile retakes. The proportions are based on tests per 1,000. For example test NWEA-394.v1 has 161 measured incidents in 5056 tests administrations, which is a rate of 32 per 1,000 or 3.2%.

Figure 4: Overall Security Incident Rates

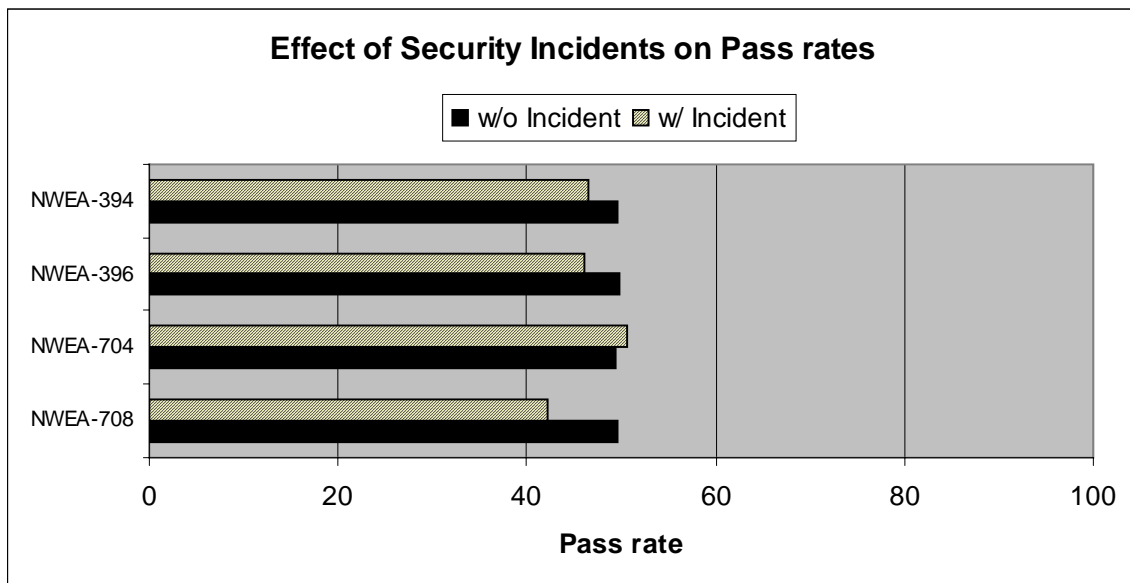


⁶ An incident is measured and counted when the statistic being compared is extreme. As such, a security incident should not be construed as confirmation of an actual security compromise. It should be interpreted as an event indicating risk to the test's security.

The effect of security incidents on the test pass rate is shown in Figure 5. This figure shows the relationship between the pass rate (a passing score was set artificially for this test at the median, because the test is used for general accountability and there was no passing score set).

The pass rate varies among the exams when comparing exams with and without security incidents across the different test forms. For example, for test NWEA-394, 46.6% of the exams that had a security incident resulted in a passing score, whereas the pass rate is higher at 49.5% for tests taken without a security incident. Normally, we have seen the opposite effect on pass rates. Having recently reviewed other educational data, two explanations for this effect are offered.

Figure 5: Effect of Security Incidents on Pass Rates

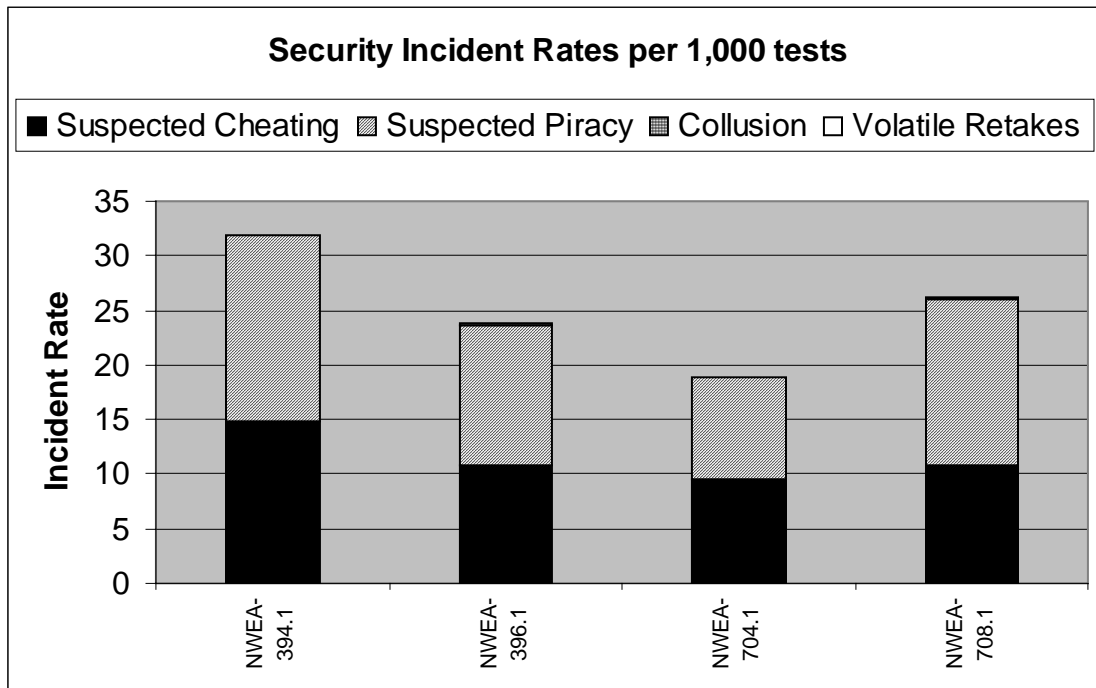


First, when students are not prepared to take the exam, high aberrance may be seen in the lower score range. It is likely that these students are receiving help to raise their scores, but without adequate preparation, the help is insufficient. A simple comparison of pass rates cannot detect this effect.

Second, the passing score selected for the analysis is not likely to be the actual passing score. The passing score for the analysis was set arbitrarily at the median. For some educational tests used for accountability purposes (e.g., NCLB) there may not be passing scores, per se. There may be cut scores associated with different proficiency levels, however, and these scores would be useful for this analysis. The passing threshold (or other classification cut scores) will nearly always be sensitive to the tail of the distribution (either low or high, depending on the targeted passing rate). Depending on the nature of the test compromise, aberrance rates will vary with the test scores. For example, if a group of “over-achieving” students have come together to “ace the test,” then aberrance will be seen at the high score levels. If test coaching is concentrated on the students in the middle of the score range then aberrance will be observed at and above the middle of the score distribution. These kinds of behavior that compromise the test may not be directly measured as an effect on the pass rate when the passing score varies or when there are multiple cut scores for multiple classifications.

Figure 6 provides the rates⁷ of security incidents per 1,000 tests for the tests. Stacked bars are used to show all the information.

Figure 6: Security Incident Rates by Type of Incident

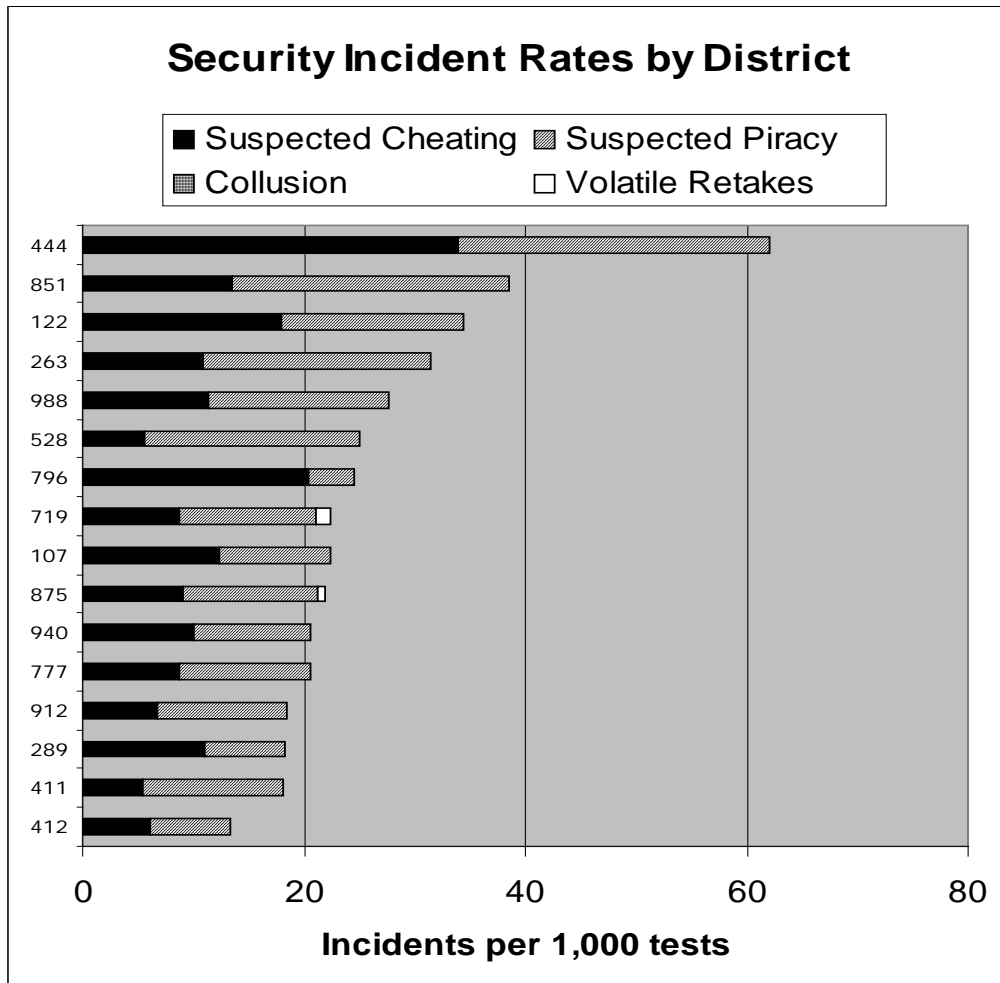


The comparison of the rates between the tests is illuminating. The greatest observed difference is between tests 394 and 704. The relative proportions between “suspected cheating” and “suspected piracy” for the exams appear to be relatively constant. This is probably a result of the CAT (Computerized Adaptive Testing) format for the exams, which will discourage cheating and aberrance.

Figure 7 presents the breakdown of security incident rates by the districts in the study. As was already noted, the two prevalent types of incidents are suspected

cheating and suspected piracy. Two of the districts had relatively high rates of incidents that would raise a flag about the validity of their test results.

Figure 7: District Overview of Security Incidents



A total of 237 students were identified as having odd response patterns or odd response latency. Forty-one of these were among the 3283 students with modified response patterns. Thus, slightly over one percent of the total number of examinees was identified as having anomalous results, and about the same percentage of the students with modified response patterns was identified. It is

not known at this time if the 41 students identified from the modified data set were among those who had the greatest number of score changes (they would be the lower scoring students who had the most items changed – 12 to 14 items). Some of the students with modified responses had as few as one or two items changed, making it virtually impossible to detect them unless their subsequent test performance would have made the anomalous responses more obvious.

Of the 237 students who were identified as having aberrant response patterns, 150 were high performing students (above the 75th percentile) whose response patterns had not been modified.

DISCUSSION

Although it is possible that the statistical procedure was identifying cheating behavior in the data, it was not able to identify the data records that had been modified to reflect students who had inappropriate access to 10% of the item pool. It might be that this manipulation, although seemingly very large, was, in reality, too subtle to be identified. It may also be that the procedures designed to work with fixed-form examinations don't work well with adaptive tests in which few students see the same items. It may also be that detection procedures designed for fixed length, linear exams (either paper and pencil or CBE) do not work well in an educational setting where we expect more variation in performance from one school to another.

The statistics that were used in this analysis are z-scores. These statistics are means of independent random variables and the Central Limit Theorem can be used to assume the statistical distributions are approximately normal. The critical value was chosen so that only 5% of the time would the maximum z-score in the sample of size 20,661 exceed the critical value. This value was set at 4.7. The alpha-level of this value is .000001, or about 1 chance in a million. Even though the test statistic is not normally distributed, the normal approximation is sufficiently close that we can be assured that the alpha-level of the statistic is very small ($<.001$). The expected number of reported cheaters by chance alone in this study would be less than 10 ($10,221 \times .001$).

Therefore we are left with a puzzle. Is the statistical procedure at fault, or is there a substantial amount of pre-knowledge already present in the unmodified data? Because the relative proportions of the detected cheaters in the known cheating set versus the unmodified data set are nearly identical: 1.25% versus 1.13%, we are left to conclude that the data forensics analysis shows no difference between the modified and unmodified data sets. Given the amount of known modification, we are left to conclude that cheating prevalence in the unmodified data is probably as large as the induced cheating prevalence in the modified data. The slightly lower rate in the modified data set could be due to the fact that items were changed only if they were answered incorrectly and no other responses were changed as would be the case in an actual testing situation.

In the unmodified data set, if the examinee answered correctly he or she was routed to a different item than if the item was answered incorrectly.

Simulation results indicate that the power of these test statistics is very low when students are armed with only 10% pre-knowledge. At an alpha-level of .001 on a 60-70 item test, the simulation results indicate power or detection rates for Caveon's best statistics are 6%. At an alpha level of .0001 in the same simulation, detection rates are approximately 2%. The simulation was not performed with extremely low alpha levels below .0001. The CAT algorithm, by virtue of adapting the test to the student's ability level, will typically lower the probability of a correct response, making pre-knowledge more difficult to detect than when the probability of a correct response is greater than 50%. This is because cheating detection relies on finding item responses that are improbable (and usually incorrect). The improbability evidence is stronger when the probability of a correct response is higher.

Given the extremely conservative testing procedure used by the data forensics analysis and the low proportion of pre-knowledge on the exam, the above results are not surprising. Cheating detection is an extremely hard problem. The difficulty is compounded by the requirement that the procedures be conservative in order to minimize false positives.

RECOMMENDATIONS

Based on the analysis and results, some recommended actions that NWEA may want to consider are as follows:

General

- Ensure that test proctoring and administration procedures are being followed.
- Perform spot audits of the testing. Concentrate efforts in districts that are showing high incident rates and suspicious security related activity.

Exam 394

- Verify that this exam is functioning as designed and that the observed instability (as seen by a pass rate jump in tandem with high degrees of aberrance) was transitory:
- Monitor aberrance and pass rates for Exam 394 or review data from the exam administered during the 2004-2005 school year to ensure that the test has not been compromised. If it has been compromised revise the exam as schedule and resources allow.

Teachers

- Reduce and deter test “coaching” by teachers.
- Coaching appears to be occurring at relatively low levels, but there are a few locations that indicate test coaching or other forms of

inappropriate examinee assistance may be taking place. Reinforce exam administration procedures in training. Also, inform local district personnel of situations that should be monitored.

REFERENCES

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum Associates.

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, April, 2001, Seattle, Washington.

Cohen, A. S., and Wollack, J. A. (in press). Test administration, scoring and reporting. To be published in *Educational Measurement*, Edition 4, 2005.

Davey, T, and Nering, M. (2002). Controlling Item Exposure and Maintaining Item Security. In Mills, C. N., Potenza, M. T., and Fremer, J. J. and Ward, W. C. (Editors). *Computer-Based Testing: Building the Foundation for Future Assessments*, Lawrence Erlbaum Associates: Mahwah, New Jersey.

Josephson, M. and Mertz, M. (2004). *Changing Cheaters: Promoting Integrity and Preventing Academic Dishonesty*. Josephson Institute of Ethics, Los Angeles, California.

McCabe, D. L. (2005). CAI Research. Center for Academic Integrity. http://www.academicintegrity.org/cai_research.asp.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., and Velasquez, R. (2004) Psychological Testing on the Internet: New Problems, Old Issues. *American Psychologist*, April, 2004, Vol 59, No. 3, 150-162

APPENDIX A

Glossary of Terms

- Aberrance Threshold:** The characterization of a test as aberrant hinges on the application of an “aberrance threshold”; a percentile on the aberrance score distribution for all tests above which differences in test-taking behavior (responses and response times) are deemed to be “significant” and indicative of test abuse.
- Aberrance Rate %:** This is the percentage of administered tests that were counted as aberrant (either the response latency aberrance statistic exceeded the threshold or the response aberrance statistic exceeded its corresponding threshold). An aberrant test exhibits response and response time values which significantly deviate from the test’s normative response model. A test is characterized as “aberrant” if its aberrance score exceeds the aberrance threshold.
- Aberrance Score:** A statistic computed by comparing observed test response and response time patterns with a model of expected response and response time patterns. Deviations from the model (abnormal test response and response times) result in a positive aberrance score.
- Alpha:** The Type I error rate that is set for the statistical tests. Because multiple tests are performed (perhaps several thousand), the thresholds must be carefully adjusted to maintain the Type I error rate. Consequently, many results which would normally be reported as significant are not indicated as significant in order to avoid inflation of the Type I error rate.
- Cheating Index:** The statistical index that measures the test of significance for the high-score aberrance rate. The null hypothesis is that the high-score aberrance rate is the same as for all other geographical units in the world (excluding the unit being tested). The index is the absolute value of the logarithm (base 10) of the p-value of the test. This allows immediate interpretation of the index in odds language. An upper-tailed test is performed. High index values indicate aberrance rates for high-score tests above and beyond world levels. High values of this index indicate elevated levels of cheating.
- High-Score Aberrance Rate:** The percent of high-score (or passed) tests that are aberrant.
- High-Score Threshold:** A percentile of the distribution of all test scores above which a test is considered to be a “high-score test.”
- Latency Aberrance Threshold:** A threshold normed against the standard normal distribution for counting whether a test is aberrant based upon the item response latency aberrance indices.
- Low-Score Aberrance Rate:** The percent of low-score (or failed) tests that are aberrant.
- Mean Score :** The average test score for all tests. The mean score of all test scores is typically the proportion of test items answered correctly, unless items scores are weighted.

- Pass Rate Index:** The statistical index that measures the test of significance for the pass rate. The null hypothesis is that the pass rate is the same as for all other geographical units in the world (excluding the unit being tested). The index is the absolute value of the logarithm (base 10) of the p-value of the test. This allows immediate interpretation of the index in odds language. A two-tailed test is performed. If the pass rate is lower than the expected value, then the index will be negative.
- Pass Rate %:** For tests where a passing standard is applied; the percentage of all tests which received a passing score.
- Piracy Index:** The statistical index that measures the test of significance for the low-score aberrance rate. The null hypothesis is that the low-score aberrance rate is the same as for all other geographical units in the world (excluding the unit being tested). The index is the absolute value of the logarithm (base 10) of the p-value of the test. This allows immediate interpretation of the index in odds language. An upper-tailed test is performed. High index values indicate aberrance rates for low-score tests above and beyond world levels. High values of this index indicate elevated levels of test piracy.
- Response Aberrance Threshold:** A threshold normed against the standard normal distribution for counting whether a test is aberrant based upon the response aberrance indices.
- Report Section:** Column in the report parameters page for the appropriate threshold level. For example, the threshold of 1.771 will be set for the world section when alpha is at .05.
- Test Site:** The location where a test was administered.
- Tests:** Count of number of tests.

A Conceptual Framework For Judging Ethical Violations And Determining Sanctions

Karen E. Banks

Wake County Public School System

In the world of educational testing, situations occasionally occur in which the ethical issues at stake and the appropriate sanctions are clearly “black and white.” Unfortunately, there are a great many times when the events and appropriate sanctions are neither black nor white, but rather fall somewhere on a

varied and shifting palette of gray. As a profession, we may someday soon reach consensus, if we have not already, about how to deal with the black and white issues, such as cheating. The challenge will be to deal with the gray ones.

INTRODUCTION

The topic of cheating is receiving a great deal of attention at present from both the media and from within the ranks of educators. Problems continue to come to light in Texas (Huff, 2005; Manzo, 2005), but Texas is not alone. For example, in the fall 2004, Mississippi found more than two dozen cases of alleged cheating at nine schools. In Nevada, testing irregularities, including student cheating and teacher misconduct, increased by more than 50 percent in 2003-04 from the previous school year (Hurst, 2004.) The state of Arizona found over 20 cases of confirmed cheating in the period of 2002-2004. The media reports go on and on (Axtman, 2005).

The NATD/NCME session of which this presentation is a part focuses largely on cheating, and this paper does touch on the issue, specifically on sanctions for cheating. Sanctions for cheating are *not* the primary focus of the paper, however, since the author believes penalties for cheating are less difficult to decide upon than are sanctions for many of the other ethical violations related to testing. In fact, the paper addresses cheating as just one type of ethical lapse, within a broader array of ethical lapses. While the paper will briefly address the

issue of what states and districts do about cheating when it happens, the focus is more on what to do about those ethical violations when someone clearly did something “wrong” but where cheating cannot be proved. Finally, the paper will try to start a conversation among the audience and the subsequent readers of the paper about how we as a profession believe educators should act in different situations. What are the criteria that will help guide us when the answers are not so clear? The criteria will have to be developed collectively, but perhaps some ideas in this paper will start the dialogue that leads to consensus.

While there is a great deal of discussion about cheating—including reasons that it happens, how to go about detecting it, and how to prevent it in the first place—many behaviors don’t reach the standard of “cheating” but they clearly violate professional ethical standards, including codified standards of various states. (Examples and hyperlinks to the state standards that are included in the reference section of this paper include the states of Florida, North Carolina, Texas, and Wisconsin.) For example, a teacher who carelessly stores secure, high stakes test materials in an unsecured location has behaved in an unethical manner according to most of the published ethical standards regarding testing. How severely should we deal with the teacher’s behavior?

The disciplinary actions taken in response to such unethical behaviors seem to vary considerably from state to state and from district to district. Deliberate cheating, when it can be proved, is probably handled more consistently than other infractions. For example, a when a teacher is found to

have given students the answers, dismissal is usually immediate, although the teacher may be allowed to resign instead. One hopes that the State Board of Education in most states would also revoke the teacher's license, although my research for this paper suggests that this does not happen consistently. State education agency attorneys contacted as background for this paper indicated that districts sometimes terminate teachers for cheating without forwarding a request to the state for license revocation, nor do all cases that come to the state's attention result in further action against a teacher that has already resigned, been terminated, or left the state.

The Need For Consistency

In the face of a general consensus about sanctions for cheating, then, this paper will argue that the problems we should be most concerned about are not only more frequent than actual cheating, but are those for which the appropriate response or sanction to an ethical violation is less clear in nature. An incident in which a teacher claims to accidentally have given students extra time to take the test, resulting in students having to be retested because of the teacher's carelessness, might result in a stern verbal reprimand in one district. Given the same situation in another school district, the wrath of parents and students that is likely to fall on the teacher due to the burden of the retesting might be considered sufficient punishment of the teacher for the error. In yet a third

district, a written reprimand that will remain in the teacher's personnel file might be considered appropriate for the same mistake.

Among the ethical principles to which most societies subscribe is the principle of fairness. If we are going to fairly address ethical behavior of school and district staff, it seems that at the least, similar actions should be punished with approximately the same level of severity, or we will not even meet the basic requirement of fairness.

While this author will not argue for the concept of national "sentencing guidelines," the paper will share what some educators in North Carolina have developed. It began after a series of incidents in the Wake County Public School System in 2003. After years of only minor ethical violations or testing irregularities among district staff members, a dozen serious incidents occurred in approximately three weeks. It was important to have some consistency across the incidents in terms of how they were handled. Telephone calls to other districts found widely varying practices in terms of how severely or leniently similar incidents had been punished. The author, who is an assistant superintendent in the district, then developed a framework that helped guide decisions about appropriate punishments for staff members who violated the N.C. Testing Code of Ethics. A few months later, the district shared this framework with testing staff from local education agencies (LEAs) across the state and received suggestions for improving it, with particular help from staff in

the neighboring Johnston County Schools. Later it was posted on an electronic bulletin board used by testing staff across the state.

Reportedly, this framework is particularly useful when dismissal and license revocation are probably not appropriate. Principals, school district Human Resource directors, and school district Testing Directors seem to have few questions about what should happen when a teacher changes a student's answer sheet. There seems to be agreement that changing even one answer on one student's answer sheet, if this unethical act can be verified, should result in the employee being fired and result in losing his or her license to teach. What about mistakes or poor judgment, however, even those with serious consequences?

ADDRESSING UNETHICAL BEHAVIOR

Education professionals – teachers, administrators, and others-- have a duty to do more than just “not cheat.” Professional guidelines concerning testing are familiar to NATD and NCME members (e.g., JCTP, 2004,) although these guidelines may not be disseminated to most building level educators. Most LEAs have a policy or code of conduct that requires ethical behavior and compliance with applicable rules and regulations. Such policies often allow for sanctions up to and including dismissal when a staff member fails to comply, but they may not be very specific. Hence, states often take steps similar to North Carolina, which codified a code of conduct around testing (North Carolina

Testing Code of Ethics, 2000.). Yet, clearly not every broken rule should lead to dismissal. Even if we agree on what is unethical behavior, how do we address it when it occurs?

TWO FACTOR MODEL

In developing the proposed framework, in discussion with other LEAs, the author settled on two factors that seemed important to consider in determining sanctions. First, was the action intentional or unintentional? Second, how serious were the consequences of the action?

In our legal system, however, which was one context for the proposed framework, the consequences do matter. If you drive while intoxicated, you will lose your driver's license. If, in the process, you injure or kill someone with your car, you will face much more extreme sanctions. This legal context is probably important as we think about sanctions for unethical testing practices because of the likelihood that an employee may seek legal remedies if sanctions are considered too severe for the specific situation. Referring to the chart on the next page, you will see examples of deliberate acts with minor consequences and those with major consequences. All deliberate acts are serious, but not all should result in dismissal of an employee or loss of a teaching license.

An ethicist might argue that intent is all that matters, and that the consequences should not be considered. If you intentionally store test materials in an unsecured location, the fact that no one stole any of the materials means that you were simply lucky. Why should your punishment be less severe than someone who committed the same infraction but who was unlucky enough to have a major theft occur?

The chart also contains what appeared to be appropriate sanctions. These will vary from state to state, based on state laws concerning personnel records, teacher licensure, and termination. However, the framework proposes what should happen, to include:

- *Corrective conference*, essentially a formal meeting in which the irregularity is discussed and the employee agrees to avoid committing such behaviors in the future
- *Letter of reprimand* in a school-based file, which is less serious than the official personnel file
- *Letter of reprimand* in the district's personnel file
- Additional personnel actions may be taken as necessary
 - Suspension with or without pay
 - Termination
 - Revocation of license (by the State Board of Education)
- Criminal or civil action

Following the chart, are some case studies. Readers are invited to consider how the incidents would have been judged in their state or district.

Severity of Consequences	Severity of Behavior		
	Negligence	Serious Negligence	Willful Intent
Moderate (Examples)	<ul style="list-style-type: none"> ▪ Room or area for test storage not secure for brief periods (e.g. 5 minutes) but all booklets accounted for and no evidence of cheating is apparent ▪ Failure to follow prescribed procedures in test training and test administrator’s manual (e.g. time limits, seating arrangements, IEP requirements,) but retesting or other measures can address the problems created. <p><i>(Corrective conference held.)</i></p>	<ul style="list-style-type: none"> ▪ Failure to provide secure, locked storage but all booklets accounted for and no evidence of cheating is apparent ▪ Failure to keep the “secure storage area: secure but all booklets accounted for and no cheating is apparent ▪ Failure to provide/attend required training ▪ Failure to follow daily check-in and check-out procedures for materials ▪ Repeated violations due to negligence <p><i>(Corrective conference held. Letter of directive in school-based file.)</i></p>	<ul style="list-style-type: none"> ▪ Knowingly signing false inventory of materials ▪ Failure to report infractions of others ▪ Allowing another person to access secure tests or test items ▪ Reviewing secure tests or test items <p><i>(Corrective conference held. Letter of reprimand in personnel file. Additional personnel actions may be taken as necessary.)</i></p>
Severe (Examples)	<ul style="list-style-type: none"> ▪ Failure to use reasonable precautions to ensure discipline and control of students before, during, and after the testing, resulting in loss of materials or other serious problems. <p><i>(Corrective conference held. Letter of directive in school-based file.)</i></p>	<ul style="list-style-type: none"> ▪ Failure to provide secure, locked storage that results in theft, cheating, or other serious consequences ▪ Failure to follow daily check-in and check-out procedures for materials that results in theft, cheating, or other serious consequences ▪ Loss of materials with no evidence of theft or cheating ▪ Failure to take corrective action once problems are known <p><i>(Corrective conference held. Letter of reprimand in personnel file. Additional personnel actions may be taken as necessary.)</i></p>	<ul style="list-style-type: none"> ▪ Repeated intentional violations ▪ Altering answer documents ▪ Providing students with answers ▪ Any criminal act, including but not limited to theft <p><i>(Serious personnel actions will be taken.)</i></p>

CASE STUDIES AND THE NEED FOR A MORE COMPLEX FRAMEWORK

The chart in the previous section has been helpful in addressing a number of problems that have arisen in some districts, but there are many grey areas left unresolved in such a chart. The first two of these cases studies will be discussed with the audience for this session and the second two merely shared. The questions posed are intended to broaden the dialogue so that we can move beyond the “two factor model” of looking at intent and consequences to clear guidance about how professional educators address ethical issues in testing?

Case Study 1

Dr. Griffin teaches geography at a military school known for its strict, no-tolerance policy on cheating. Among questions on a recent test, Dr. Griffin asked students to name the capital cities of 100 countries listed. Sitting in the third row were two students. The young woman was, he had come to recognize, one of the strongest students in the class. Beside her was a young man who has been struggling. When Dr. Griffin graded the tests, he noted that the young woman had misspelled six capital cities, although she had the cities with the correct countries. The young man sitting next to her spelled the

same six cities wrong, using exactly the same misspellings as the young woman.

Dr. Griffin did not see the male student cheat, but the similarity in error patterns was a concern. Nor could Dr. Griffin prove that, if cheating did occur, that the young man copied from the woman, rather than the other way around. Finally, he could not say whether the two students had colluded in some way.

His school expects him to report such incidents as “potential cheating.” Dr. Griffin understands that to discuss this issue with his department head would cause the entire situation to, as he said, “Go nuclear.” An all out investigation would follow and the young man would probably be permanently expelled based on the limited evidence available. He feels that this situation is not clear enough for such severe actions against the student.

What is Dr. Griffin’s ethical responsibility to report this situation under these circumstances? What other actions might he need to take, if he chooses not to report it? If he does not report it and this failure to report discovered, what actions should the school take against him? Against the students?

Case Study 2

Amy Ward is a high school student in the International Baccalaureate (I.B.) program. She is a good student, primarily as a function of her hard

work, and has done well in her classes. She typically finishes her exams well within the time allotted, and with high quality responses. In reviewing one of Amy's I.B. exams, the instructor – Mr. Smith – noticed that Amy had done a really good job on the short essay questions on pages 1-8 and pages 10-14. Amy had somehow completely skipped page 9, however, and it was not clear she had even seen the page since there were no marks on it.

When Mr. Smith questioned Amy, she became very distraught. She indicated that although she had a cold that day, she thought she had answered every question and that she must have skipped that page by mistake because she certainly had plenty of time to have finished more questions. She even posited that somehow the page might have been flipped accidentally when she got up to get a tissue or had been stuck to the other page.

Mr. Smith faced a dilemma. Should he send in Amy's test to be graded just as it is or allow Amy to complete page 9 of the booklet under timed testing conditions? This is a high stakes test for students, the results of which affect their diplomas, college entrance, and credits earned. It is unlikely that a student could pass an exam when skipping an entire short essay question. Amy will not receive an I.B. diploma if she does not pass this test. Mr. Smith feels that Amy's test booklet, as it currently exists, does not contain a true

estimate of her achievement and he believes her description of events to be plausible.

Case Study 3

Miss Baker is a first-year teacher who was responsible for administering the North Carolina End of Grade (EOG) tests to her third grade class. She attended all the training provided to test administrators at her school and read the administration manuals. Due to the large number of rules listed in the manual, it was hard to remember all of them. For example, special education students could use highlighters, but other students could not. Diabetic students could have a snack during testing, but other students could not. Other rules pertained to the types of assistance that could and could not be given to students during the test.

As she walked around the room early on the first morning of testing, Miss Baker noticed that Jason had marked all C's on the first eight multiple choice questions. Miss Baker said quietly, "Jason, I'd like you to look at those eight questions and try again." This statement was not part of the script or list of statements test administrators were directed or allowed to use. The test proctor assigned to monitor the test administration in Miss Baker's class later reported the incident, indicating that she felt it was clear that Miss Baker was telling Jason that his answers were wrong and to change them.

What type of action should have been taken against the teacher? Does it matter that the teacher is in her first year and may not have completely understood the rules?

Case Study 4

Mrs. Jolly was an experienced teacher in a large middle school and part of a four-teacher team. Her classroom was typically used for science classes. On the day of the mathematics achievement test, she was responsible for administering the test to 25 students. Some of the math word problems involved using formulas to calculate volume, while others asked students to convert temperatures from Celsius to Fahrenheit. The bulletin board on the wall was not covered and had charts that talked about metric units of volume and about Fahrenheit and Celsius temperature scales. Once Mrs. Jolly realized that the mathematics test had word problems related to the content of the bulletin board, she immediately notified the school test coordinator.

While students reported that they did not use the bulletin board information to answer the test questions, all students were required to be re-tested with an alternate form of the mathematics test. Parents were upset that the teacher's error resulted in their students being subjected to another morning of testing. The dilemma concerned the level of discipline appropriate for the teacher. What type of action should have been taken? Does it matter

that the teacher is an experienced teacher? That she reported her own mistake?

SUMMARY

In an ideal world, the incidents described above would be dealt with in the same way in each classroom and across all states and school districts. An ethical violation that leads to dismissal in one school or district should not lead solely to a verbal reprimand in the district next door, no matter how stern that verbal reprimand might be. While the framework proposed is a beginning, it leaves some unanswered questions about many of the problems faced by states and local school districts. Continued dialogue is needed – to the point of consensus, if possible – if we are to have a fair system in place for dealing with the ethical problems that affect our testing programs.

REFERENCES

- Axtman, K. (2005, January 11). When tests' cheaters are the teachers. *Christian Science Monitor*, <http://www.csmonitor.com/2005/0111/p01s03-ussc.html>
- Florida State Board of Education, (1994, October.)Rule 6A-10.042, FAC. <http://www.firn.edu/doe/rules/6a-106.htm#6A-10.042>
- Hoff, D. J. (2003) New York teachers caught cheating on state tests. *Education Week*, 23 (10), p. 27.

Hoff, D. J. (2005, January 12). Texas officials to investigate allegations of cheating. *Education Week*, 24 (18), p.21.

Hurst, M. D. (2004) Nevada report reveals spike in test irregularities. *Education Week*, 24 (6), pp.19, 22.

Joint Committee on Testing Practices. (2004) *Code of fair testing practices in education*. Washington D.C., Joint Committee on Testing Practices.

Manzo, K.K. (2005) Houston inspector finds cheating on state tests. *Education Week*, 24 (37) p.4.

North Carolina State Board of Education (2002, August). *Testing code of ethics*. <http://www.ncpublicschools.org/docs/accountability/testing/policies/testcode080100.pdf>

Texas Administrative Code, Chapter 101. (2001, November). <http://www.tea.state.tx.us/rules/tac/chapter101/ch101c.html>

Wisconsin Department of Public Instruction, (2004.) *Guidelines for appropriate test security*. http://dpi.wi.gov/oea/doc/wkce_final04_guidlms.rtf

Twenty Years of Cheating: From Sneaky Students to Shifty Systems

Joseph O'Reilly
Mesa Public Schools

Welcome to the 20th anniversary of a NATD Symposium on cheating! It is interesting timing that we are discussing this issue today. On Thursday several teachers in my district were given sanctions for using actual test items from a district social studies exam for test preparation. And as we are

speaking three more who are in leadership positions are finding out that their worlds are changing. It appears that someone served to help write the test and brought his knowledge as well as other things back to the other teachers. And the saddest part is that we didn't notice it because the school out performed its peers, in fact it slightly underperformed them even with this extra help. Which is something Jim Impara found in one sample.

So this topic has been very much on my mind lately.

And it has been a hot topic for NATDers for many years. In 2001 we had a symposium that also focused on cheating, but the discussion centered on the individual cheater. The good news is that cheating was found to be relatively rare – that is what test directors said in a survey. It is what Kentucky found, 62 instances in 463,360 tests and what Michael Kean experienced, with two instances in 8½ years in Philadelphia.

But it is important. As Glynn Ligon said at a similar symposium twenty years ago,

*Teachers cheat when they administer standardized tests to students.
Not all teachers.
Not even very many of them.
But enough to make cheating a major concern to all of us who
use test data for decision making.*

And those last five words are one key reason why it is important to us as test directors. We are responsible for the quality of our tests. We are tasked with ensuring that the tests yield valid and useful information and that we

will guard against threats to the accuracy of the testing information. And cheating is a threat.

What is different about the session today is that we are not focusing on the individual cheater – the prevalence or the methods like cameras in ballcaps, but on the system in which cheating occurs. And that is important. The cheating I started out mentioning was a systemic problem that was pervasive and sanctioned, explicitly or implicitly by people up and down the chain. It wasn't just the one teacher who checked wrong answers on a student's paper and said try again [as often as three out of the four choices] because as she said "*how else was I going to get my bonus?*" It was the system. As more sanctions are placed on schools and more of us use bonuses, the system, in a department, in a school, in a district, will become a bigger issue. So it is good that we are talking about this today.

We started with Greg Cizek's paper on the systemic influences on cheating. He makes a point that looking at just the individual is not enough, we have to look at the system. And he is exactly right.

One of the issues he brings up is having a common definition of cheating. One thing that is not stressed is that one needs a common definition before the cheating happens.

Although when defining cheating I would prefer to paraphrase Justice Potter Stewart's definition of pornography – I may not be able to define it but I

know it when I see it, our district has a 19 page 'Directives for Testing' that lays out what is good practice and what is not allowed and teachers sign a statement each year that they have read it. While those caught often claim that *"you don't specifically state that you can't _____"* (whatever it is they were caught doing), having defined parameters earlier make it much easier to judge behavior.

But I think Cizek is right on target in saying there is a need to raise awareness of everyone in the system of that common definition of what is and what is not acceptable. Even with our guide and teacher sign off, I don't think we do enough to make it explicit and salient every year.

Cizek also calls for measures for identifying cheating. We saw one such method in the Impara et. al. paper. But in 2001 Cizek wrote, and I quote, *"Statistical analyses should be triggered by some other factor (e.g, observation). None of the statistical approaches should be used as a screening tool to mine data for possible anomalies."* I would like to hear Cizek and Impara discuss this issue.

And that is a good segue into the next paper on detecting cheating. The attitude here is, "cheating happens, identify it." And the authors show some good ways to identify it, although they have a ways to go in improving the system because it does not quite work as well as it needs to yet. As they say at the end *"cheating detection is a very hard problem."*

One concern I have about this approach as someone who deals with individual cheaters is the difference between statistical truth and legal truth. In the Caveon approach the authors are looking at probabilities and there will always be some chance that an individual did not cheat. But when we are dealing with someone or a system accused of cheating we are dealing with legal truth – did this individual, in this situation really cheat beyond a shadow of a doubt? The approach we saw today can shine the light on an issue, but cannot alone prove it.

The value of this approach, however, is that it show us where to look. In 2001 Kentucky reported 62 instances of suspected cheating out of a half million tests. If Ligon’s earlier quote is correct, do you think the state identified all the cases? Probably not. This approach tells us where to focus our efforts, where to look for those additional cases.

Even if the statistical approach used just led to asking the right people questions about how they gave a test, you would do a lot to raise the perceived risk of being caught to offset any benefit of cheating. We found that out when the word got out that the district did ‘erasure analysis.’ It looked official because we dummied up an official looking printout. Behind that Potemkin printout was not a sophisticated software package but an individual looking through the bubble sheets when they came in and flagging those

classes with a lot of erasures for further scrutiny. But the number of tests with excessive erasures, although infrequent already, fell precipitously.

Although the authors did not succeed in identifying all the cheating on the CAT tests, they did raise some important issues. First, as we move to computerized testing and computerized adaptive testing, we have to look closely at our data for anomalies. And approaches like this one that flags outlying events makes it easier for us to do that.

Second, the anomalies raise the issue of what are proper, and effective, test preparation practices. The software can identify high aberrance among low scoring students – those who received some specific coaching but who do not do well overall on the test. Take the teachers I started talking about who gave the kids the test items in advance and the students still did not do well. Overall, the software may have highlighted the situation for us sooner.

Third, I never really thought how CAT makes it harder to cheat but it also may make it harder for us to catch anomalies that trigger investigations of possible cheating. If students never get a chance to get those easy items they didn't have in advance wrong, we may not catch them.

Finally, it raises the question – do we really want to know? How many state or large district testing programs use an approach such as this? We spend hundreds of millions of dollars on creating and giving tests in this

country, but very little on ensuring their validity and accuracy. And even less on resources to follow up on anomalies.

So after noting that cheating has a systemic component and there are ways to detect it what do we do? Cizek suggests an honor code. Karen Banks takes it to a more detailed and specific level – having a common framework for judging ethical violations and determining sanctions.

As we saw, this isn't an easy issue. In fact, judging individual situations is quite hard. But, as Cizek mentioned, it is very important to clearly communicate to teachers.

And that is why we need just what Banks is proposing be done in advance and widely distributed. This tells people what is not acceptable and what will happen if you cross a line. It makes the system fair, or at least equitable. And it makes it explicit.

Banks says that she is not 'arguing for national sentencing guidelines,' and I would agree we don't want mandatory sentences. But there is a need to spread her idea of consistency across districts. We would be well served by model guidelines of best practices.

Creating model guidelines of how different behaviors are defined as cheating and dealt with would be very useful. Especially if it was agreed to by the professional groups representing teachers and administrators. Set up like JCTP, it could be a joint effort that is led by NATD to define what we as

professionals consider cheating and what we think should happen when it occurs. We just wouldn't want to call it the Joint Committee on Cheating Practices, because someone may misinterpret the group's mission. And who better to lead such an effort than someone who is retiring with all the time in the world and half her professional life ahead of her?

One concern I have in setting out the consequences is how they will impact the reporting of cheating. Teachers now report most cheating to us. If the consequences are too severe, I think we may see less reporting. As it is, I know of a principal who told his teachers not to report an incident to the district so he didn't have to deal with the hassle of an investigation. Imagine how people would feel if they were going to lose their teaching license. We need to have buy-in by teachers and administrators that these guidelines are the best professional practices in their area. That is why a joint committee is so important.

So, to sum up, in 1985 and 2001 NATD discussed the individual as a cheater and now in 2005 we are discussing how to deal with cheating from a systemic perspective – with how the system condones it, how it can be caught in a systemic way and how we need to deal with it consistently across levels and systems.

I hope the next time we get together on this topic we are here to discuss how we as professional organizations propose to address the issue of

cheating. Or, better yet, how well our common efforts to minimize cheating are working. And I hope we can say that after putting cheating behaviors under a microscope it still is an infrequent occurrence.