

An Overview of Issues Concerning Cheating on Large-Scale Tests

Prepared by:

Gregory J. Cizek
Associate Professor of Measurement and Evaluation
110 Peabody Hall
School of Education, CB 3500
University of North Carolina
Chapel Hill, NC 27599-3500

Paper presented at the annual meeting of the National Council on Measurement in Education, April 2001,
Seattle, WA.

An Overview of Issues Concerning Cheating on Large-Scale Tests

Cheating undermines integrity and fairness at all levels. It leads to weak life performance. It undermines the merit basis of our society. Cheating is an issue that should concern every citizen of this country. (Cole, 1998, p. A-24)

Sound testing practices and the high quality information that can result, are helpful to those who have oversight, responsibility, or interest in American education. From a broader perspective, sound testing programs yield benefits to society at large (Mehrens & Cizek, 2001). To the extent that tests provide high quality information, better judgments about individual students can result. Test data also provide the grist for pursuing well-reasoned courses of action in terms of recommendations for improving policies and practices and evaluating reforms.

Just as clearly, factors that attenuate the validity of tests or degrade the usefulness of information yielded by them represent threats to sound decision making. It can be said that the primary role played by those in the field of psychometrics is what might be called "data quality-control specialist"--helping to ensure that tests yield the kind of valid and useful information that they were designed to produce. One aspect of data quality-control is a professional vigilance about threats to the accuracy and dependability of test information.

To a great degree, modern testing theory and practice have evolved to potently address many of the threats. For example, validity theory has been advanced through the work of Kane (1992), Messick (1989), and others. Generalizability theory (Brennan, 1992) provides a sophisticated new way of examining the dependability of test scores. The state of the art in setting passing scores has advanced more in the past decade than perhaps over any other period (Cizek, 2001). Computerization has made automated test assembly and administration as common in high-stakes testing contexts as the #2 pencil (Luecht, 1998). The

degree and breadth of these changes is witnessed by the recent, extensive revision to the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999).

Despite these advances, areas for increased vigilance remain. One perplexing challenge is presented by the problem of cheating on tests. As the use of tests for informing decisions increases, and as the stakes associated with test performance rise, the problem of cheating has become more prevalent. One need not consult psychometric journals for evidence of the phenomenon; even a casual reading of the local newspaper demonstrates that the incidence and extent of cheating is on the rise. A few examples:

* According to an Associated Press report, "At least 52 teachers from five states cheated on their competency tests by paying \$1000 bribes to exam supervisors for extra time and help with the answers" (Payne, 2000, p. A-7).

* A front-page article in the *Detroit News* reported on called for "a full investigation into the Dearborn [Michigan] Police Department's promotional tests, following a prosecutor's report that concludes someone in the city altered test results" (Merx, 2000, p. A-1).

* Spanning just the last three years, stories in the *New York Times* have documented cheating on tests assessing a wide range of vocations, knowledge, and skills; tests of citizenship, truck driving, stock brokering, and others have made the headlines (see, Seelye, 1998; Toy, 1999; Sullivan, 1997). A prominent series of stories has chronicled the recent investigation of widespread cheating on student achievement tests by New York City educators (see Stancik, 1999).

Some Background on Cheating on Examinations

Although cheating on tests has increasingly been given attention in the popular media, it has received only modest attention by those actually involved in the field of testing. However, the topic is still one that is

frequently swept under the carpet or ignored. To begin a consideration of cheating, it is appropriate to first define the kinds of testing situations in which cheating can occur. Then, in subsequent sections of this article, I will present a summary of some methods for detecting cheating and suggestions for how cheating can be prevented.

Contexts and Varieties of Cheating

In this article, I limit the domain of cheating to that which occurs on large-scale educational achievement tests such as the kinds of pupil proficiency testing increasingly mandated by individual states. These tests are typically administered to students at prescribed grade levels, (say, at grades 4, 8, and 12) to measure student progress at key junctures in their education (such as at exit from elementary, middle, and high school). Some states choose a off-the-shelf test such as the *Iowa Tests of Basic Skills* or *Stanford Achievement Test* for monitoring purposes. These tests are designed to assess a set of objectives that would be broadly common across the country. Other states contract with test development companies to produce custom tests that are carefully aligned to the state's unique curriculum framework.

Regardless of the source of the tests, the content covered by them is most often limited to basic subject areas such as reading, writing, mathematics, science, and civics. Usually, the format of the tests is a combination of the familiar multiple-choice questions and so-called "constructed-response" items in which the student, for example, solves a mathematics problem, writes an essay, or conducts a science experiment. A pupil's performance on the test may be reported as simply passing or failing, or a system of ordered descriptors of performance may be used, such as *Advanced*, *Proficient*, *Nearly Proficient*, and *Beginning*. Increasingly, consequences for students are being attached to performance on the tests. For example, in some states, passage of a 4th grade reading test may be required for promotion to the 5th grade, or passing the 12th grade test may be required in order to receive a high school diploma. Tests with weighty consequences such as these are referred to as "high-stakes" tests.

There are many forms of cheating that can occur on such tests. In general, cheating can be defined as

any action that violates the rules for administering a test, any behavior that gives an examinee an unfair advantage over other examinees, or any action on the part of an examinee or test administrator that decreases the accuracy of the intended inferences arising from the examinee's test score or performance. A person may not actually take a test himself or herself, but may use another person (called a "confederate") to take the test in his or her place. An examinee may use unauthorized materials such as a "cheat sheet," or take advantage of the testing situation by, for example, requesting testing accommodations that are not necessary. One method of cheating on large-scale examinations involves taking advantage of time zone differences when tests are to be administered at sites across the U.S. at the same time. Using this method, examinees at one test site in, say, New York City, complete an examination in time to telephone other examinees in, for example, Los Angeles, to communicate the content of the test in advance. (To my knowledge, there are no organizations that address this method, despite the fact that it has been documented to have been used extensively to cheat on tests such as the Graduate Management Admissions Test (GMAT), the Test of English as a Foreign Language (TOEFL), and the Graduate Record Examination (GRE), and the fact that simply staggering test administration times would address the problem. Overall, these actions represent broad categories of behaviors; a more extensive and specific list of cheating methods is provided by Cizek (1999).

For their part, test developers usually produce carefully scripted directions for administering their tests and provide clear guidelines for what kinds of behaviors on the part of examinees are permissible and which are not. Acceptable and unacceptable behaviors are sometimes formalized in states' administrative codes or statutes. Numerous professional organizations have published statements on the inappropriateness of cheating. Some of the most explicit statements regarding cheating are found in the aforementioned *Standards for Educational and Psychological Testing*. Among other things, the *Standards* indicate those involved in testing programs should:

- * protect the security of tests (Standard 11.7);
- * inform examinees that it is inappropriate for them to have someone else take the test for them, disclose secure test materials, or engage in any other form of cheating (Standard 8.7);
- * ensure that individuals who administer and score tests are proficient in administration procedures and understand the importance of adhering to directions provided by the test developer (Standard 13.10);

- * ensure that test preparation activities and materials provided to students will not adversely affect the validity of test score inferences (Standard 13.11); and
- * maintain the integrity of test results by eliminating practices designed to raise test scores without improving students' real knowledge, skills, or abilities in the area tested (Standard 15.9).

In summary, there has not been a dissemination problem regarding what constitutes integrity in testing, and what constitutes cheating on tests. Virtually everyone involved in testing knows how to administer (and take) tests that yield credible, accurate results. Unfortunately, mere knowledge about what constitutes cheating is not enough.

Why Cheating is a Problem

Validity is the single greatest concern in any testing situation. The concept refers to the accuracy of the interpretations about examinees based on their test scores. Phrased in only slightly more technical terms, validity is the degree to which evidence supports the inferences made about a person's knowledge, skill, or ability based on his or her observed performance. By definition, inferences are based upon a less-than-ideal amount of information, such as on a sample of a person's knowledge or skill obtained via a test. Because it is often too costly or impractical to gather more information, inferences must be based on samples of behavior. Consequently, it is necessary to consider the accuracy of inferences based on the available evidence (e.g., test performance); that is, to consider validity. This idea of validity as accuracy-of-inferences and sufficiency-of-evidence are central in modern psychometric theory and are the foundation of professionally defensible testing practices. Any factor that attenuates the ability to make accurate inferences from the sample of performance threatens validity and jeopardizes the meaningfulness of conclusions about the test taker. When cheating occurs, inaccurate inferences result.

Who Cheats, How Much, and Why?

Test takers cheat. They let others cheat. Test administrators and proctors cheat. Although hard data on the frequency of cheating is difficult to come by, there is some evidence. The data we do have is of two

types: results of research studies on cheating (most often surveys), and anecdotal reports that arise via newspaper and broadcast media outlets. Both sources of evidence have limitations. Surveys always suffer from some degree of inaccuracy; the concerns are heightened when the questions center on sensitive or illegal behaviors. Anecdotal reports are sometimes exaggerated or prove to be false. Despite these limitations, the incidence of such anecdotes has increased dramatically, and there is enough credible research evidence accumulating to conclude that the problem of educators cheating on tests is occurring and is occurring more frequently. Summarizing across studies, a range of 3-5% is a reasonable estimate of the percentage of examinees who engage in cheating on any particular occasion. Anecdotal reports confirm that it is not only test takers who are cheating. As was documented in the cases described previously involving bribes paid to proctors and in the far-reaching investigation in New York City schools, those who give tests are also engaging in the behavior with surprising frequency.

It is easiest to comprehend examinees' motivations for cheating. They want a license to practice in their chosen profession, higher grades, opportunities for advancement, issuance of a credential, and so on. Sometimes examinees allow other test takers to cheat. One study by Davis, et al. (1992) conducted with college students, examined the reasons why they would do so, with the top reasons shown in Table 1.

Insert Table 1 about here.

Cheating by those who give tests is only slightly more difficult to understand. Basic motivations provided by bribes or other rewards certainly exist. Those who teach courses, direct residency programs, or oversee education and training organizations have an interest in promoting strong performance on the part of their students. Finally, numerous research studies have documented that the majority of high school and college graduates cheated on tests in their own academic careers; because so much cheating of that cheating

went undetected and unpunished, and because they can easily put themselves in the position of examinees desperate to pass a test, those who give tests may often be tempted to turn a blind eye to cheating.

Cheating as an Irregular Event

A conclusion that cheating has occurred on a test can only be made after a careful examination of evidence. Usually, such an investigation begins following what is initially termed a "testing irregularity." When tests are administered, events that are out-of-the-ordinary can occur. These events may be within or beyond the control of those administering and/or those taking tests and, until causal attributions can be confidently asserted, cannot be interpreted as cheating. A first step in detecting cheating is to have in place a set of procedures for observing and documenting irregularities. Examples of irregularities could include:

- 1) a fire alarm activation that required evacuation of a building during a testing session. Ordinarily, this event would be beyond the control of test administrators, but the event could increase student anxiety, reduce students' ability to attend to test materials on their return to the testing session, etc. If this occurred, students' performances on the test may not represent their true levels of knowledge, skill, or ability; that is, the students' proficiency levels would be underestimated.
- 2) permitting examinees to have additional time to complete a test beyond the limits indicated. This event would ordinarily be within the control of test administrators. If this occurred, examinees' performances on the test may again not represent their true levels of knowledge, skill, or ability, though in this case, students' proficiency levels would likely be overestimated.
- 3) repeated, sustained glancing by one examinee at the answer sheet of an adjacent examinee.

Two fundamental questions arise when a testing irregularities occurs. One question concerns the

likelihood of the event. Unusual or out-of-the ordinary occurrences happen all the time; however, some events are less likely than others. The more unlikely an event is to occur, the more our curiosity is piqued. For very rare events, such as winning a Super Lottery or being struck by lightning, the probability of their occurrence is often of great interest.

The second question centers on explanations for unusual events. For example, airplane crashes are very rare; the intense interest in understanding the cause of that rare occurrence can linger for months, even years following the event. Our interest is particularly keen in understanding what role, if any, human intervention may have played in the event. Purely random events occur all of the time and they can be readily accepted as such. For example, in a fair lottery, numbers are selected randomly and those who do not hold the winning number can (usually) accept the randomness of that event. On the other hand, it would not be tolerable if human intervention or manipulation of the Lottery number selection tilted the process in favor of certain numbers or gave an *a priori* advantage to certain individuals. This type of human intervention changes our characterization (and acceptability) of the process from random to fraudulent.

The responsibilities of those who administer tests mentioned in the previous section are particularly germane to this point. When suspicion exists that testing irregularities may have occurred as a result of human intervention--either through negligence, deviation from prescribed testing practices, or intentional manipulation of circumstances, testing conditions, or results--then our sense of ethical behavior and fairness is violated as are, in many cases, legal or administrative guidelines.

Methods for Evaluating Testing Irregularities

There are two general categories of methods for investigating and evaluating testing irregularities: judgmental and statistical. As the label suggests, judgmental methods rely more heavily on subjective human judgments. For example, a student might enlist the aid of a confederate to take an examination in his or her place. Human judgment is involved in detecting and responding to this irregularity when the proctors for the examination scrutinize photo identification before permitting examinees to take the test. Judgment is also

involved when handwriting samples from the student are compared with those of the confederate to make a determination of whose handwriting appears on the test materials.

Statistical methods can be used to estimate the likelihood of events, such as anomalous or unusual test results. Some events have probabilities associated with them that are very small. For example, the first-year National Hockey League team, the Columbus Blue Jackets, are estimated to have only a 1 in 500 chance ($p = .002$) of winning the 2001 Stanley Cup. Those odds are actually fairly good compared to chances of being struck by lightning are (1 in 709,260 or $p = .00000141$); the chances of dying from a lightning strike are even less (1 in 2,794,493 or $p = 000000358$). Worse yet are the chances of correctly picking six numbers out of 49 in a lottery (1 in 14,000,000 or $p = .000000071$).

All of the p -values mentioned in the preceding paragraph refer to extremely small probabilities. In fact, the examples selected illustrate occurrences that could be considered nearly impossible. What is the threshold that should cause us to consider an event as being so unlikely due to chance that we are compelled to consider other potential causes? In the social sciences, the standard probability level associated with statistical significance (that is, the p -value at which scientists come to conclusions and/or make decisions about human behavior) is $p < .05$.

Of course, highly unlikely events *can* occur. However, as mentioned previously, we ordinarily become suspicious when highly improbable results occur, and we are led to conclude that simple chance should be ruled out as a plausible explanation. Should the Blue Jackets win the Stanley Cup, the circumstances surrounding such an upset in the expected course of events would likely lead to calls for an investigation of any irregularities in that sporting contest. Similarly, unusual results can occur on tests. For example, two examinees seated next to each other may and taking a 200-item multiple choice licensure examination, may answer 146 items correctly. Further, they may choose the same incorrect options for the 54 items they answered incorrectly. Statistical methods for detecting cheating on tests answer the simple question: "How likely is it that these examinees would, *by chance alone*, have produced the same response patterns?" If the answer to that question suggests that the events were *not* very likely due simply to chance,

then investigations into plausible alternative explanations begins.

It is important to note, however, that statistical methods do not obviate the need for human judgment. Even once test results are shown to be highly unlikely, human rationality must be invoked to come to any conclusions about whether alternative causes represent more plausible explanations for the results; that is, there still exists a need to make subjective interpretations about whether the unlikely events represent cheating.

Triggering Investigations of Testing Irregularities

It is not enough to ascertain that a testing irregularity was an improbable event. As mentioned previously, improbable events do occur. The probability of obtaining a score of 20 out of 20 through blind guessing on a test comprised of True/False items would be $p = .000000954$ --a nearly impossible event. However, other factors would ordinarily alter our interpretation of that probability. For example, if an examinee did not answer the 20 items through blind guessing, but used his or her knowledge of the content being tested to make more informed answer choices, then the probability would be substantially reduced. Further, if the test were an extremely easy one, and if the examinee were highly knowledgeable, then the probability of obtaining a score of 20 out of 20 could approach $p = 1.0$. Thus, to evaluate the probability of an occurrence, we must bring ancillary information to bear.

One increasingly essential source of supplemental information is referred to as a “trigger.” In large testing programs such as the SAT, for example, many people obtain scores that are highly unusual (e.g., a total score of 1600). Such performance would not arouse suspicion of an irregularity if the student had taken the test previously and obtained a 1560, if the student had a high school GPA of 4.0, was class valedictorian at a college preparatory school, and so on. On the other hand, such performance *would* arouse suspicion, for example, if the examinee’s previous performance had been a 470, if a fellow student reported that the examinee had access to the SAT test questions in advance, or if a test proctor observed the examinee copying from a test-taker of extremely high ability who was seated nearby. Each of these latter situations involves

what is called a trigger--additional information that suggests further investigation of the irregularity is warranted.

In cases cheating is suspected, statistical evaluations of test results are usually not appropriate in the absence of a trigger. However, the presence of a trigger necessarily changes our interpretations of the likelihood that results were obtained fairly. Suppose, for example, the 20-item true/false test described earlier involved simple multiplication facts. It would be highly unlikely for a three-year-old child to obtain a raw score of 20. Statistical estimates of the probability of the event would be very small, but the small probability would not necessarily lead to an allegation that the result was improper. However, if an observer of the child during the test reported that he or she saw the child's parent whispering something in the child's ear immediately prior to the child answering each of the questions, that information--a trigger--would suggest that the unusually unlikely event be regarded with a heightened level of suspicion, and that other, plausible explanations for the child's amazing performance be investigated. Common triggers for conducting statistical investigations of alleged cheating include such things as observations by a proctor of unusual examinee behavior during an examination, or anonymous "tips" or reports that an examinee had access to a secure examination materials prior to the administration.

Of course, triggers usually involve human judgment and, as such, can be fallible. The extensive literature in the field of criminology speaks definitively about the unreliability of eyewitness testimony. An act of inference occurs when a proctor observes one examinee cheating by looking at another examinees' answer sheet. Objectively, the behavior can also be interpreted as an examinee innocently averting his or her gaze temporarily to gain relief from intense concentration on the task at hand.

Statistical Tools

A number of statistical tools exist to help in the detection of possible cheating and provide quantification of the probability that an irregularity can be attributed to chance. One commercially-available software program has been developed, and is currently available from Assessment Systems Corporation

(ASC) of St. Paul, Minnesota. The program is called *Scrutiny!* and it can be run on a typical personal computer. *Scrutiny!* uses an approach to identifying copying called "error similarity analysis" or ESA--a method which, unfortunately, has not received strong recommendation in the professional literature. One review (Frary, 1993) concluded that the ESA method: 1) fails to utilize information from correct response similarity; 2) fails to consider total test performance of examinees; and 3) does not take into account the attractiveness of wrong options selected in common. Bay (1995) found that ESA was the least effective index for detecting copying of the three methods she compared.

Despite this technical weakness, *Scrutiny!* and its accompanying documentation provide a thorough introduction to the logic of detection and sound advice regarding appropriate cautions for interpretation and use of the results. Additionally, the software is easy to use, is compatible with many common input file formats, permits the user to enter a seating plan (which can be used in conjunction with identified examinee seat locations to corroborate the suspicion of cheating) and produces traditional test summary statistics as well as the probability statistics of interest regarding potential answer copying.

Other statistical indices for detecting potential copying exist, though they are not yet available in commercially available software packages. Two such indices, g_2 , developed by Frary et al. (1977) and \hat{u} , developed by Wollack (1997) are technically superior to the ESA method.¹ These procedures offer more power to detect copying, while safeguarding against over-identification of copying (i.e., Type I errors or false positives), and can be used with relatively small sample sizes (i.e., around 200 examinees). Unlike the ESA and other methods that rely only on common numbers of errors which can bias results when overall ability is not accounted for, g_2 and \hat{u} incorporate information from common right answers and differential probabilities of selection of incorrect options.

Although *Scrutiny!* or other methods may provide a defensible way of producing evidence to support a suspicion of cheating, it is important to restate that 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. A recent court decision involving an the Association of Social Work Boards

(ASWB) examination program provides an illustration. According to an article in the *ASWB Association News* (Atkinson, 2000), several examinees who had taken the February 1995 administration of the ASWB examination had their scores invalidated and were refused the issuance of licences. These actions were the result of analyses of their test scores which "revealed statistical abnormalities" (p. 9). In litigation, it was noted that "there did not appear to be any on-site problems" or reports of irregularities when the test was administered, although an "administrator for the social work board had received a telephone call indicating that certain individuals had copies of the exam prior to its administration" (p. 9). Both the circuit court and appeals court decided in favor of the examinees, noting that there was a lack of evidence to suggest why the examinees were investigated for possible cheating in the first place. It appears that, the telephone call notwithstanding, no triggering event was found such as would justify the consideration of statistical evidence.

The Particular Problem of Educator Cheating on Tests

The testing director of a large city school district summarized the problem: "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).

One need only search the internet, look at an issue of a national magazine, or skim a newspaper, to confirm that many educators are attempting to circumvent the testing, monitoring, or accountability systems. Stories of cheating abound, and the methods are numerous, ranging from subtle coaching to overt manipulation. A *US News and World Report* article described a case in Ohio, where one educator is accused of physically moving a student's pencil-holding hand to the correct answer on a multiple-choice question (Kleiner, 2000). A recent *Washington Post* story announced the resignation of a Potomac, Maryland principal who stepped down amidst charges that she "was sitting in the [class]room, going through test booklets and calling students up to change or elaborate on answers" (Schulte, 2000). A colleague of mine in educational testing tells the 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 his choice of words like "salutation" wasn't done randomly, but was done with the vocabulary section of the state-mandated, norm-referenced test in hand.

I found out about one of the most blatant forms of educator cheating over a decade ago at an evening reception following a conference for school district superintendents in one midwestern state. I happened upon a conversation among several superintendents who, with cocktails in hand, were chuckling and winking about how their quality control procedures for student testing involved "pre-screening the kids' answer sheets for stray marks." What was so funny--I found out later from one of the superintendents--was that "stray marks includes things like wrong answers." Wink. Apparently, the practice continues. Another recent article describes how 11 school districts in Texas are being called to account for an unusually high number of erasures on that state's test (Johnston & Galley, 1999).

Most cheating is probably not as overt. More subtle forms of cheating are undoubtedly more frequent, but still serve to degrade the meaning of test results and confidence in education systems. The more subtle kinds of cheating occur when a teacher prods a student to review his or her answer: "Why don't you take another look at what you wrote down for number 17." Some of those who give tests cheat by proxy, when they fail to monitor test taking and effectively encourage cheating on the part of students. Educator cheating also occurs when they fail to include all students who would be eligible to take a test, as is the case when a teacher reminds certain students who are likely to obtain low scores on a test that it would be OK for them to be absent on the day of the test. The *Education Week* article by Johnston & Galley (1999) described a sophisticated variation of this kind of cheating in which incorrect student identification numbers were apparently purposefully entered on the answer sheets of low-scoring students, which had the effect of kicking those answer sheets out of the scoring process and raising the school's average performance. In other states which require that an absent student be recorded a score of zero (which would lower a school's average performance) all students are encouraged to attend on the day of testing, but some are afforded "testing

disability accommodations” such as an individual aid, reader, or other assistance not usually a part of the student’s educational experience.

Perhaps the most visible report of cheating by educators involved teachers and principals in the New York City school system. An exhaustive study of cheating was conducted by Edward Stancik, Special Commissioner of Investigation for the New York City School District. The study found that cheating by 12 educators was “so egregious that their employment must be terminated and they should be barred from future work with the [Board of Education]” (Stancik & Brenner, 1999, p. 63) The report named another 40 educators who were recommended for disciplinary action 35 of whom engaged in actions judged serious enough to warrant potential termination. Examples of the cheating Stancik identified included those of a principal, who during a test “walked around the room and pointed out [to the students] incorrect choices, saying either “That’s wrong” or “Do that one over” (p. 2). According to Stancik’s investigation, 4th-grade students at another school reported that their teacher, Teresa Czarnowski, helped them cheat by correcting their answers in advance. Stancik reported: “According to one boy, who is indicative of those we interviewed, after he finished the test on the separate sheet [of scrap paper], he gave it to Czarnowski who checked his choices and marked an X on the scrap next to his wrong answers. Then she returned the paper to the student who corrected his responses and, finally, he transferred his selections to the official bubble form” (p. 11). Overall, the report concluded that there had been “extensive cheating by educators” that the school district had “known about the problem for years” and that “educators were no held fully liable for their misconduct” (p. 60). The public release of the initial report brought greater attention to the problem. According to a follow-up report issued in May 2000 by the investigators’ office:

“Almost immediately, our intake unit was busy with new complaints of wrongdoing committed by Board of Education employees during the testing process. Then in February 2000, while we were conducting investigations into those allegations, students took the State English Language Assessment (ELA) examination and reports of suspicious behavior and writing in test booklets again

poured into our office.... Once again we found proctors who gave answers to students, alerted them to wrong responses, and changed student choices after the exam was turned in. Moreover, this investigation uncovered new methods of misconduct, including prepping children for the third day of the ELA exam by using the actual test material. Finally, our investigations continued to be impeded by delay in the reporting of testing allegations to this office.” (Stancik, 2000, p. 1)

The follow-up report named another 10 educators who had engaged in seriously inappropriate behaviors during testing in New York City. For many of the educators named, the cheating was so blatant--for example, writing answers to test questions on the chalk board--that immediate termination of employment was recommended.²

Research on Educator Cheating

The most common avenue of research does not ask educators directly about whether they engage in what have come to be referred to euphemistically as “inappropriate test administration practices” though a few studies have done so. Usually, educators have been polled regarding their general perceptions of cheating in their schools. One such study asked 3rd, 6th, 8th, and 10th grade teachers in North Carolina to report how frequently they had witnessed certain inappropriate practices. Overall, 35% of the teachers said they had observed cheating, by either personally engaging in inappropriate practices or being aware of unethical actions of others. (The teachers in this study reported that their colleagues engaged in the behaviors from two to ten times more frequently than they had personally.) The behaviors included giving extra time on timed tests, changing students' answers on their answer sheets, suggesting answers to students, and directly teaching specific portions of a test. More flagrant examples included the case of students being given dictionaries and thesauruses by teachers for their use on a state mandated writing test. One teacher revealed that she checked students answer sheets "to be sure that her students answered as they had been taught." Other teachers reported more subtle strategies such as "a nod of approval, a smile, and calling attention to a given answer"

were effective at enhancing students' performance (Gay, 1990). Another study of teachers drawn from two large school districts found that 31.5% of the teachers surveyed reported spending two or more weeks giving students old forms of standardized tests for practice (Shepard & Dougherty, 1991).

In a study initiated to investigate suspected cheating in the Chicago Public Schools, a total of 40 schools were included, 17 as "control" schools and 23 "suspect" schools which exhibited irregularities in the performances of their 7th and 8th grade students on the *Iowa Tests of Basic Skills* (ITBS). The irregularities consisted of unusual patterns of score increases in previous years, unnecessarily large orders of blank answer sheets for the test, and high percentages of erasures on students' answer sheets. The researchers readministered the ITBS under more controlled conditions and found that, even accounting for the reduced level of motivation students would have had on the retesting, "clearly the suspect schools did much worse on the retest than the comparison schools" and concluded that "it's possible that we may have underestimated the extent of cheating at some schools" (Perlman, 1985, pp. 4-5). A study of cheating in the Memphis school district revealed extensive cheating on the *California Achievement Test*, including one case in which a teacher displayed correctly filled-in answer sheets on the walls of her classroom (Toch & Wagner, 1992).

Educators' Perceptions of Cheating

Perhaps the most troubling stream of research on cheating concerns the attitudes of educators toward cheating. Generally, there appears to be a growing indifference on the part of educators toward the behavior and even an increasing sense that cheating is a justifiable response to externally-mandated tests.

Several attempts have been made to investigate educators' perceptions of cheating. In one study, 74 pre-service teachers were asked to indicate how appropriate they believed certain behaviors to be. Only 1.4% thought that either changing answers on a student's answer sheet or giving hints or clues during testing were appropriate, and only 2.7% agreed that allowing more time than allotted for a test was acceptable. However, 8.1% thought that practicing on actual test items was okay, 23.4% believed rephrasing or rewording questions to be acceptable, and 37.6% judged practice on an alternate test form to be appropriate (Kher-

Durlabhji & Lacina-Gifford, 1992).

The beliefs of pre-service teachers appear to translate into actual practices when they enter the classroom. A large sample of 3rd, 5th, and 6th grade teachers in two school districts was asked to describe the extent to which they believed specific cheating behaviors were practiced by teachers in their schools. On the positive side, their responses (shown in Table 2) indicated that for all of the behaviors listed but one, a majority of respondents said that they occurred rarely or never. Equally noticeable, however, is that a wide range of behaviors was reported as occurring “frequently” or “often” by, in some cases, 15% or more of respondents. A second observation that leaps from Table 2 is the remarkable extent to which teachers report that they have “no idea how often this occurs” (Shepard & Dougherty, 1991).

Table 2 about here.

Another survey examined perceptions about two specific kinds of “test preparation” practices: having students practice for a state-mandated, norm-referenced test using another form of the same test, or having students practice on the actual test that would be used. The survey polled six groups of educators, including teachers and administrators drawn from schools in the midwestern U.S., and teachers, principals, superintendents, and school board members from California. The results, shown in Table 3, reveal fairly broad acceptance of these behaviors, even among board members.

Table 3 about here.

Though not attempted here (or elsewhere to my knowledge), the costs of cheating probably could be measured in dollars and cents. What cannot be measured are the effects of educator cheating at more

fundamental levels. For example, when students learn that their teachers or principals cheat, what is the effect of this kind of role modeling? While fallen professional athletes might be able to say, “Don’t look at me as a role model, I am just an athlete doing a job,” educators cannot: a significant aspect of their job *is* the modeling of appropriate social and ethical behavior. Also, how might educator cheating affect students’ attitude toward tests or their motivation to excel? How might it affect their attitude toward education, their trust or cynicism with respect to other institutions, or their propensity to cheat in other contexts?

Preventing Cheating

What can be done to deter cheating? Fortunately, many things. As a starting point, it important to note that bringing the issue of cheating forward as a topic for discussion is likely to increase awareness of the problem on the part of those who give tests and those who take tests. It is important to heighten sensitivity about a validity threat heretofore virtually ignored. From the broadest perspective, it may be useful to entirely reconceptualize testing so that successful test performance can be more consistently and directly linked to student effort and effective instruction, and so that unsuccessful performance is accompanied by sufficient diagnostic information about students’ strengths and weaknesses. As a result of identification and remediation of those weaknesses, we advance the persepective that obtaining accurate test results is more beneficial to all concerned than cheating (Cizek, 1999, chap. 11).

There are also numerous, more pragmatic steps that can be taken. The following list should provide a start. Of the following, some are focussed on test givers; some on test takers; some apply to both.

- 1) Get the word out. It has been said that we more often stand in need of being reminded than we do of education. As mentioned previously, nearly all testing programs provide documentation describing appropriate test administration procedures, state regulations define legal conduct for test administrators, and professional associations have produced documents to guide sound testing practice. Nonetheless, reports of those accused of cheating on tests are often accompanied by the

protestation that they did not know the behavior was wrong. If only as a reminder and to heighten awareness, every implementation of high-stakes tests should be accompanied by dissemination of clear guidelines regarding permissible and impermissible behaviors. Such reminders should be clearly-worded, pilot-tested, distributed, and signed by all who handle testing materials, including test site supervisors, proctors, and examinees.

2) Decrease reliance on easily-corruptible test formats. Changes in test development practice can reduce the potential for some methods of cheating. For instance, it is more difficult for one examinee to copy another examinee's answer to an essay question, case analysis, or other constructed-response format than it is to copy a bubbled-in response or provide the key to a multiple-choice item. It must be recognized, however, that such changes require tradeoffs in terms of efficiency and scoring costs.

3) Limit the amount of testing. It is probably a truism that limiting the amount of testing will decrease the amount of cheating. As many states continue to expand their pupil proficiency testing programs as a primary mechanism for accountability, opportunities for cheating are expanded. There have been two, common, reactionary responses to the predictable increase in cheating. One reaction is the demand that large-scale testing for accountability be abandoned. For example, the September 22, 1000 issue of the *Congressional Quarterly* contained an essay by Monte Neill, the executive director of a group critical of testing, who argued the "pro" position on the question "Should high-stakes tests be abolished in order to reduce cheating?" (Neill, 2000). In the same issue, education writer Alfie Kohn is noted as one of several critics who "have seized on cheating as just another in a long list of reasons to abandon [standardized] tests." According to Kohn, "The real cheating going on in education reform is by those who are cheating students out of an education by turning schools into giant test-prep centers" (quoted in Koch, 2000, p. 759).

Related to the first reaction are demands that responsibility for judging student achievement be located more within individual educators' sphere of professional responsibility. A second reaction is that testing for accountability rely more heavily on constructed-response type item formats which, ostensibly, would be less prone to corruption. For instance, it is argued that it is more difficult to forge or coach a student's answer to an essay question or a science experiment than to alter a bubbled-in response or provide the key to a multiple-choice item.

The difficulty with these first-blush reactions is that they fail to fully address the core issues. As I have argued elsewhere, the genesis of high-stakes pupil testing in the 1970s was made inevitable because of poor decision making--or at least perceived poor decision making--and the resulting search for alternatives (identifying reference omitted). It was during the tumultuous 1970s that complaints of some business and industry leaders began to receive broad public currency: "We are getting high school graduates who have a diploma, but can't read or write!" As Popham observed at the time: "Minimum competency testing programs ... have been installed in so many states as a way of halting what is perceived as a continuing devaluation of the high school diploma" (1978, p. 297). The clear public perception was that the gatekeepers were leaving the gates wide open. Perhaps a widespread misunderstanding of the relationship between self-esteem and achievement was to blame.

Understandably, educators wanted all students to achieve and all to have the personal esteem associated with those accomplishments. But assigning higher grades to heighten self-esteem and stimulate accomplishment too often had neither effect. The sense that grades weren't all they were cracked up to be wound its way from business and industry leaders' lips to policy makers' pens.

As the line of reasoning went, if the gatekeepers of the 1970s weren't keeping the gates as conscientiously as the public had hoped, then important decisions about students should be remanded

to rely on passing one or more common tests. Thus, the obvious error in current calls to return to the past is that such a strategy only puts American education back in a place that caused accountability tests to be introduced in the first place. Moreover, though current tests have been shown to be susceptible to cheating, the solution of returning to measures and procedures that were demonstrably even more easily manipulated is unthinkable.

What should be considered is limiting the amount of testing for accountability. We must remember that there is a distinction between instruction and evaluation. It is obvious that not all tests are done for the purposes of evaluation. Equally true, however, is that not all tests--especially those designed for purposes of decision making--must have instructional value. Once their purpose has been clarified, the scope of mandated accountability tests, the time required for their administration, and the opportunities for cheating can be minimized.

4) Revise "Truth in Testing" laws. States with so-called "truth in testing" laws should reconsider the relative benefits such laws. These laws often require that the questions on tests used to monitor student achievement or for accountability purposes be publicly disclosed following administration of a test. Despite their good intention, the unforeseen consequence of such laws has been an increase in educators' use of previous versions of tests for classroom practice, resulting in further narrowing of instruction. Additionally, the economic costs to states with such laws has been staggering, brought about by the need to develop entirely new monitoring instruments one or more times each year.

5) Audit test security procedures. Those with oversight for testing programs can incorporate operational changes--many of which require only modest changes in current procedures--that can have a cumulative positive effect on reducing cheating. Many of these are not new, and many may already be in place. However, a regular "security audit" to review procedures is desirable. Common

security measures include shrink-wrapping, numbering, and bar-coding of test materials to deter unauthorized access and to permit tracing the path that the materials take. Other simple steps can easily be added, such as delaying delivery of testing materials until just prior to test administration, and, once delivered, requiring that materials be maintained securely by a named person responsible for the materials.

6) Improve test administration conditions. Increased attention must be paid to one of the weakest links in the security chain: proctoring. Too often, the qualifications for supervising or proctoring examinations are only faintly spelled out, the training provided is minimal if any, and no incentives exist to heighten their vigilance or pursue instances of cheating. For all testing contexts, proper training must include instruction on methods examinees use to cheat, how to approach a test taker regarding suspicions of inappropriate behavior without unduly disrupting other examinees or inducing anxiety in those who are not cheating. In the context of large-scale testing, training should include effective procedures for documenting on-site testing irregularities.

7) Use available statistical tools. Finally, we recall that statistical detection methods should not be used as screens for statistically unusual response patterns. Nonetheless, research has demonstrated that informing examinees that detection software will be used can dramatically reduce the incidence of cheating. One study by Bellazza and Bellazza (1989) showed a reduction from approximately 5% to 1% in the amount of cheating on college-level management course examinations. If a detection program may be used to provide supplemental evidence following a triggering event, it makes sense to inform examinees that detection software may be used.

8) Provide penalties for cheating and change the system of investigation. In conjunction with limiting opportunities for cheating, procedures for investigating cheating and penalties for educator

cheating must be dramatically revised. Currently, many tests are administered behind closed classroom door with little independent oversight; there are strong disincentives for educational personnel to report cheating; and in most jurisdictions, the responsibility for investigating cheating involves personnel at the school or district level and agencies such as boards of education with an inherent conflict of interest when it comes to ferreting out inappropriately high, apparent, student achievement.

Revised procedures should include: random sampling and oversight of test sites; increased protections for “whistle-blowers;” more streamlined procedures and stiffer penalties including permanent disqualification from teaching within a state and more coordinated sharing of information regarding educators who have had their licenses revoked; and delegation of responsibility for investigation cheating to an independent authority.

9) Implement honor codes. Because honor codes have been shown to reduce the incidence of cheating in other contexts, their use in licensure and certification testing should be examined. Honor codes require examinees to pledge to abide by a set of standards, including eschewing cheating themselves and obligating themselves to report cheating by others. Requiring examinees to sign such a pledge prior to taking an examination may work in credentialing settings as well.

Conclusions

Overall, the evidence on the problem of cheating on tests is in. Cheating is occurring with increasing frequency. It is fair to conclude that the problem will not disappear. The problem must be addressed, however, in order to ensure the integrity, fairness, and validity of test results. As a beginning step, those who

have oversight of testing programs should make themselves aware of the myriad ways cheating can occur, including cheating by examinees and test administration staff who may aid examinees in cheating.

Additionally, those responsible for testing programs should address how they can help to reduce cheating, and should pursue courses that foster even greater levels of public protection and professional responsibility for the citizens and associations they serve.

References

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Atkinson, D. (2000, August). Testimony tests test. *ASWB Association News*, p. 9, 11.
- Bay, M. L. G. (1995, April). *Detection of cheating on multiple-choice examinations*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Bellazza, F. S., & Bellazza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology*, *16*(3), 151-155.
- Brennan, R. L. (1992). Generalizability theory [NCME instructional module]. *Educational Measurement: Issues and Practice*, *11*(4), 27-34.
- 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). Conjectures on the rise and fall of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3-17). Mahwah, NJ: Lawrence Erlbaum.
- Cole, N. (1998, November 9). Teen cheating hurts all. *USA Today*, p. A-24.
- Davis, S. F., Grover, C. A., Becker, A. H., & McGregor, L. N. (1992). Academic dishonesty: Prevalence, determinants, techniques, and punishments. *Teaching of Psychology*, *19*(1), 16-20.
- Frary, R. B. (1993). Statistical detection of multiple-choice answer copying: Review and commentary. *Applied Measurement in Education*, *6*(2), 153-165.
- Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, *2*, 235-256.
- Gay, G. H. (1990). Standardized tests: Irregularities in administering of tests affect test results. *Journal of Instructional Psychology*, *17*(2), 93-103.

- Johnston, R. C., & Galley, M. (1999, April 14). Austin district charged with test tampering. *Education Week*, p. 3.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.
- Kher-Durlabhji, N., & Lacina-Gifford, L. J. (1992, April). Quest for test success: Preservice teachers' views of high stakes tests. Paper presented at the annual meeting of the Mid-South Educational Research Association, Knoxville, TN. (ERIC Document Reproduction Service No. ED 353 338)
- Kleiner, C. (2000, June 12). Test case: Now the principal's cheating. *U. S. News and World Report*.
- Koch, K. (2000). Cheating in schools. *Congressional Quarterly*, 10(32), p. 759.
- 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).
- Luecht, R. M. (1998). Testing and measurement issues: Automated test assembly in the era of computerized testing. *CLEAR Exam Review*, 9(2), 19-22.
- Mehrens, W. A., & Cizek, G. J. (2001). Standard setting and the public good: Benefits accrued and anticipated. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 477-485). Mahwah, NJ: Lawrence Erlbaum.
- Merx, K. (2000, August 11). Cop test altered in Dearborn. *The Detroit News*, p. A-1.
- Messick, S. A. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement, third edition* (pp. 13-104). New York: Macmillan.
- Neill, M. (2000). Should high-stakes tests be abolished in order to reduce cheating? *Congressional Quarterly*, 10(32), p. 761.
- Payne, P. (2000, August 18). Officials say 52 teachers paid \$1,000 to pass competency tests. *The [Schenectady, NY] Daily Gazette*, p. A-7.

Perlman, C. L. (1985, March). Results of a citywide testing program audit in Chicago. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 263 212), pp. 4-5.

Popham, W. J. (1978). As always, provocative. *Journal of Educational Measurement*, 15, 297-300.

Popham, W. J. (1991, April). Defensible/undefensible instructional preparation for high-stakes achievement tests. Presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Schulte, B. (2000, June 1). School allegedly cheated on tests. *Washington Post*, p. A-1.

Seelye, K. Q. (1998, January 28). 20 charged with helping 13,000 cheat on test for citizenship. *New York Times*, p. A-1.

Shepard, L. A., & Dougherty, K. C. (1991). Effects of high-stakes testing on instruction. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 337 468)

Stancik, E. F. (2000, May 2). Correspondence to Harold O. Levy, Chancellor of New York City Public Schools, pp. 1-2.

Stancik, E. F., & Brenner, R. M. (1999). *Cheating the children: Educator misconduct on standardized tests*. New York: Office of the Special Commissioner of Investigation for the New York City School District.

Sullivan, J. (1997, January 9). 53 charged in brokers' testing fraud. *New York Times*, p. A-7.

Thacher Associates. (2000). *Setting the record straight: Anatomy of a failed investigation*. New York: Author.

Toch, T., & Wagner, B. (1992, April 27). Schools for scandal. *U.S. News and World Report*, pp. 66-72.

Toy, V. S. (1999, November 23). Drivers' test scheme reveals secret decoder watchbands. *New York*

Times, p. B-2.

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, 21, 307-320.

Table 1

Top Reasons for Letting Other Students Copy During an Examination

8. Just to do it. I didn't like the teacher, and I knew if I got caught nothing would happen.
7. I knew they studied and knew the material, but test taking was really difficult.
6. No particular reason. It doesn't bother me because I probably got it wrong and so will they.
5. Because they might let me cheat off them sometime.
4. She was damn good looking.
3. I wouldn't want them to be mad at me.
2. I knew they needed to do good in order to pass the class. I felt sorry for them.
1. He was bigger than me.

Adapted from Davis, et al. (1992)

Table 2

Teacher beliefs about inappropriate test administration practices

Question: To what extent do you believe these are practiced by teachers in your school?

<u>Behavior</u>	<u>Percent of respondents</u>				
	<u>Never</u>	<u>Rarely</u>	<u>Often</u>	<u>Frequently</u>	<u>No Idea</u>
1. Providing hints on correct answers	28.5	20.8	16.9	5.8	28.0
2. Giving students more time than test directions permit	38.0	19.7	15.2	4.4	22.7
3. Reading questions to students that they are supposed to read themselves	38.8	22.2	11.9	2.2	24.9
4. Answering questions about test content	43.2	20.5	8.9	2.8	24.7
5. Changing answers on a student's answer sheet 58.4	7.8	5.5	0.6	27.7	
6. Rephrasing questions during testing	36.3	20.8	16.1	1.9	24.9
7. Not administering the test to students students who would have trouble with it	50.7	15.8	7.5	5.8	20.2
8. Encouraging students who would have trouble on the test to be absent on test day	60.1	10.8	5.5	1.9	21.6
9. Practicing items from the test itself	54.6	12.5	8.0	3.3	21.6
10. Giving students answers to test questions	56.8	11.6	6.4	1.9	23.3
11. Giving practice on highly similar passages as those in the test	24.9	15.8	20.5	19.7	19.1

From Shepard and Dougherty (1991)

Table 3

Teacher and administrator beliefs about inappropriate test administration practices

	Percent of respondents considering the practice to be appropriate					
	Midwest		California			
	Teachers	Administrators	Teachers	Principals	Supts.	Board Mbrs.
<u>Behavior</u>						
Student practice with previous test form	34	47	57	25	60	68
Student practice with current test form	14	17	36	6	17	21

From Popham (1991)

Notes

-
1. The developers of the indices g_2 and \bar{u} have made programs available for calculating those indices. Readers wishing to obtain a program for calculating g_2 should send a request via email to Robert Frary at fraryrb@prodigy.net; users interested in a program to calculate \bar{u} should send an email request to James Wollack at jwollack@facstaff.wisc.edu.
 2. The Stancik report itself has not been without controversy. An investigation and report on Stancik's findings commissioned by the teachers' union (United Federation of Teachers) and conducted by Thacher and Associates (2000) was highly critical of the methods employed in the initial investigation. The report concludes that the original investigation may have incorrectly identified some educators as having engaged in inappropriate practices.