DOCUMENT RESUME

ED 480 061 CG 032 634

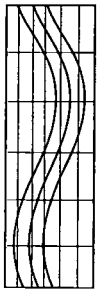
AUTHOR	Cizek, Gregory J.
TITLE	Educational Testing Integrity: Why Educators and Students Cheat and How To Prevent It.
PUB DATE	2003-08-00
NOTE	26p.; In: Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators; see CG 032 608.
PUB TYPE	Information Analyses (070)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Cheating; *Educational Testing; Integrity; *Prevention; *Student Evaluation; Test Interpretation; *Testing Problems

ABSTRACT

Sound testing practices and the high-quality information that can result are helpful to those who have oversight, responsibility, or interest in American education. To the extent that tests provide high-quality information, they form the basis for making accurate judgments about individual students. It is equally true, however, that factors which attenuate the validity of tests or degrade the usefulness of the information they yield represent threats to sound decision making. This chapter addresses both student cheating and educator cheating, and recommended strategies for preventing cheating. (Contains 38 references and 1 table.) (GCP)

Educational Testing Integrity: Why Educators and Students Cheat and How to Prevent It

By
Gregory J. Cizek



Chapter 26
Educational Testing Integrity
Why Educators and Students Cheat and How to Prevent It

Gregory J. Cizek

> *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 benefit society at large (Mehrens & Cizek, 2001). To the extent that tests provide high-quality information, they form the basis for making accurate judgments about individual students. 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.

It is equally true, however, that factors which attenuate the validity of tests or degrade the usefulness of the information they yield represent threats to sound decision making. Those in the field of psychometrics are what might be called "data quality-control specialists" who help to ensure that tests yield the kind of valid and useful information 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 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 sophisticated new ways of examining the dependability of test scores. Computerization has made automated test assembly and administration as common in high-stakes testing contexts as the No. 2 pencil (Luecht, 1998). The degree and breadth of these changes are witnessed by the recent a more

extensive and specific list of cheating methods used by test takers. Student cheating is not the only concern, however. Those who are responsible for administering tests can also act in ways that destroy the accuracy of test result interpretation, and examples of educator cheating will be provided later in this chapter.

Why Cheating Is a Problem

Validity is the single greatest concern in any testing situation. The concept refers to the accuracy of the interpretations made about examinees based on their test scores. Phrased in 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 generally 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.

Guidelines Regarding Cheating

There is an abundance of information to guide test takers and test administrators in how to avoid inappropriate testing practices. For their part, test developers usually produce carefully scripted directions for administering their tests and provide clear guidelines as to which kinds of behaviors on the part of examinees and educators are permissible and which are not. Acceptable and unacceptable behaviors are sometimes formalized in state administrative codes or statutes; one example is found in the State of Ohio Revised Code (see Amended Senate Bill 230, Ohio Revised Code, 3319.151, 1996). Numerous professional organizations have published statements on cheating (see, e.g., National Association of Test Directors, n.d.). Some of the most explicit statements regarding cheating are found in the aforementioned

Standards for Educational and Psychological Testing (AERA et al., 1999). Among other things, the *Standards* indicate that 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).

Despite these admonitions regarding cheating, not all communication about cheating is clear. For example, the same test publisher that produces a test administration manual with explicit guidelines regarding proper test administration and security procedures might also publish test preparation materials that bear a strong resemblance to actual tests. Moreover, guidelines for appropriate administration can vary from test to test, with one publisher permitting a teacher to clarify a test question for a student and another publisher proscribing the same behavior.

Although some ambiguities will always exist regarding whether a particular action constitutes cheating, there has not generally been a dissemination problem regarding what constitutes integrity in testing or cheating on tests. Virtually everyone involved in testing knows how to administer (and take) tests that yield credible, accurate results. Unfortunately, mere knowledge about what constitutes cheating is not enough.

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 are

difficult to come by, two types of data exist: 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, particularly when the questions center on sensitive or illegal behaviors. Anecdotal reports are sometimes exaggerated or prove to be false. Despite these limitations, reports of cheating are surfacing with increasing regularity, and enough credible evidence has accumulated to conclude that the problem of educators cheating on tests is increasing.

Summarizing several studies, Bellezza & Bellezza (1989) speculate that 5 percent may be a reasonable estimate of the percentage of test takers who engage in cheating on any particular occasion. And, though the frequency of educator cheating is surely small, the previously mentioned accounts of bribes paid to proctors, the far-reaching investigation in New York City schools, and other reports suggest that those who give tests are also engaging in the behavior with increasing frequency.

Examinees' motivations for cheating are easiest to comprehend. They want high grades, a license to practice in their chosen profession, opportunities for advancement, issuance of a credential, or other payoffs. Sometimes examinees allow other test takers to cheat. Davis and colleagues (1992) conducted a study of college students, examining why they would allow others to cheat; the most frequently cited reasons follow:

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

Cheating on the part of those who give tests is only slightly more difficult to understand. Teachers and principals have professional pride at stake and, increasingly, the potential for personal reward or sanction under school accountability systems. Those who direct medical

residency programs or oversee education and training organizations have an interest in promoting strong performance on the part of their students. Numerous studies have documented that the majority of high school and college graduates have cheated on tests in their own academic careers. Because so much of that cheating went undetected and unpunished, and because they can easily put themselves in the position of examinees desperate to pass a test, those who give tests may often be tempted to turn a blind eye to cheating.

A Different Way of Thinking about Cheating

The conclusion that cheating has occurred on a test can be made only after a careful examination of evidence. Such an investigation usually begins following what is initially termed a "testing irregularity." When tests are administered, events that are out of the ordinary can occur. Such an event may be within or beyond the control of those administering or those taking tests. Until causal attributions can be confidently asserted, the event cannot be interpreted as cheating. Examples of irregularities could include these:

- a fire alarm 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, or have other consequences. 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.
- permitting examinees to have additional time to complete a test beyond the limits prescribed. This event would ordinarily be within the control of test administrators. If this occurred, examinees' performances on the test again may not represent their true levels of knowledge, skill, or ability, though in this case students' proficiency levels would likely be overestimated.
- repeated, sustained glancing by one examinee at the answer sheet of an adjacent examinee.

Two fundamental questions arise when a testing irregularity occurs. One concerns the likelihood of the event. Unusual occurrences are not infrequent, but some events are less likely than others. The less

likely an event is to occur, the more our curiosity is piqued. The rarer an event is—such as winning a super lottery or being struck by lightning—the greater our interest in the event usually is.

The second question centers on explanations for unusual events. For example, airplane crashes are rare; an 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 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 tilted the process in favor of certain numbers or gave a priori advantage to certain individuals. This type of human intervention changes our characterization (and acceptance) of the process from random to fraudulent.

The responsibilities of those who administer tests are particularly germane to this point. When we suspect that testing irregularities may have occurred as a result of human intervention—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. At minimum, a first step in addressing the problem of cheating is to establish and ensure broad familiarity with a set of procedures for observing and documenting irregularities.

Assessing the Possibility of Cheating

There are two general categories of methods for investigating and evaluating the potential that cheating has occurred: judgmental and statistical. As the label suggests, judgmental methods rely more heavily on subjective human interpretations. For example, a student might enlist the aid of a confederate to take the SAT 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 in comparing handwriting samples from the student 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 very small probabilities associated with them. For example, according to gambling experts, the first-year National Hockey League team the Columbus Blue Jackets was estimated to have only a 1 in 500 chance ($p = .002$) of winning the 2002 Stanley Cup. Those odds are actually fairly good in comparison to the chance of being struck by lightning (1 in 709,260, or $p = .00000141$); the chance of dying from a lightning strike are even less, estimated at 1 in 2,794,493, or $p = 000000358$. Worse yet are the odds of correctly picking 6 numbers out of 49 in a lottery (1 in 14,000,000, or $p = .000000071$).

All the p values mentioned in the preceding paragraph represent extremely small probabilities. In fact, the examples illustrate occurrences that could be considered nearly impossible. But at what threshold should we consider an event as being so unlikely to have occurred by 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 or make decisions about human behavior) is $p < .05$.

Of course, highly unlikely events *can* occur, but we ordinarily become suspicious when they do, and we are led to conclude that simple chance should be ruled out as a plausible explanation. If the Blue Jackets were to win the Stanley Cup, such an upset would likely lead to calls for an investigation to rule out any irregularities in that sporting contest. Similarly, unusual results can occur on tests. For example, two examinees seated next to each other during a 200-item multiple-choice licensure examination may each answer the same 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.

Triggers for Investigations of Testing Irregularities

It is not enough to ascertain that a testing irregularity was an improbable event, because 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 used his or her knowledge of the content being tested to make informed answer choices or educated guesses, then the probability of scoring 20 out of 20 would be substantially increased. Further, if the test were an easy one, and the examinee 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 highly unusual scores (e.g., a total score of 1600). Such performance would not arouse suspicions of irregularity if that student had taken the test previously and obtained a 1560, had a high school GPA of 4.0, was class valedictorian at a college preparatory school, and the like. On the other hand, such performance would arouse suspicion if, for example, the examinee's previous score 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 nearby test taker of extremely high ability. Each of these situations involves what is called a trigger: additional information that suggests further investigation of the irregularity is warranted.

In cases where cheating is suspected, statistical evaluations of test results are usually not appropriate in the absence of a trigger. The presence of a trigger, however, necessarily changes our interpretation 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 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. If, however, an observer during the test reported that she saw the child's parent whispering in the child's ear immediately prior to the child answering each question, 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 should 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, anonymous tips that a student had unauthorized prior access to a test, and reports by one teacher that another teacher gave students extra time to complete a state-mandated examination.

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 (see, e.g., Loftus, 1979). An act of inference occurs when a proctor observes one examinee apparently looking at another examinee's 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 detect possible cheating and to provide quantification of the probability that an irregularity can be attributed to chance. Only one commercially available software program exists. The program, called *Scrutiny!,* can be run on a typical personal computer. Unfortunately, *Scrutiny!* has not received strong recommendation in the professional literature (see Bay, 1995; Frary, 1993). Statistical procedures for detecting copying that are technically superior to that used by *Scrutiny!* exist; however, they are not yet commercially available in software packages (see Frary, Tideman, & Watts, 1977; Wollack, 1997). These procedures offer more power to detect true copying while safeguarding against overidentification, and they can be used with relatively small sample sizes.

Although statistical 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 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 licenses. These actions were the result of analyses of their test scores that "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 and appeals courts decided in favor of the examinees, noting that there was a lack of evidence to justify the examinees being investigated for possible cheating in the first place. It appears, the telephone call notwithstanding, that no triggering event was found to 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 of educator cheating: "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 a national magazine, or skim a newspaper to confirm that many educators are attempting to circumvent the testing, monitoring, and accountability systems. Stories of cheating abound, and the methods are numerous, ranging from subtle coaching to overt manipulation. A *U.S. 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 a principal who would begin the announcements each morning with a greeting via the school 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, students learned the meanings of words like *salutation* from the principal's daily announcements; they probably never learned that his choice of such words was not random, but was made with the vocabulary section of the state-mandated, norm-referenced test in hand.

I found out about a particularly blatant form of educator cheating more than 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 state-mandated student testing involved "prescreening 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 were called to account for an unusually high number of erasures on that state's test (Johnston & Galley, 1999).

Most cheating is probably not this overt. More subtle forms of cheating are undoubtedly more frequent, but they still serve to degrade the meaning of test results and confidence in education systems. 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, by failing to proctor tests conscientiously, thereby effectively encouraging cheating on the part of students. Cheating also occurs when educators fail to include all students who would be eligible to take a test, as might happen when a teacher reminds certain students who are likely to score poorly on a test that they are permitted to be absent on the day of the test. The *Education Week* article by Johnston & Galley (1999) describes 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. This had the effect of kicking those answer sheets out of the scoring process and inflating the school's average performance. Another form of cheating involves affording a student inappropriate or unnecessary testing disability accommodations such as an individual aide, reader, or other assistance not usually a part of the student's educational experience.

Perhaps the most prominent report of educator cheating involved teachers and principals in the New York City school system. Edward Stancik, special commissioner of investigation for the New York City School District, conducted an exhaustive study of cheating. His 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 recommended another 40 educators for disciplinary action, 35 of whom engaged in actions judged serious enough to warrant potential termination. Examples of the cheating Stancik identified included 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, fourth-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 not 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 behavior during testing in New York City. Many of the educators had cheated so blatantly—for example, by writing answers to test questions on the chalkboard—that immediate termination of employment was recommended.[1]

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 percent of the teachers said they had observed cheating, in terms of 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 answers on students' 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 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).

A study initiated to investigate suspected cheating in the Chicago Public Schools included a total of 40 schools, 17 "control" schools and 23 "suspect" schools that exhibited irregularities in the performance of their seventh- and eighth-grade students on the Iowa Tests of Basic Skills. 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 Iowa Tests 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 educators' attitudes toward cheating. Generally, educators appear to be

growing increasingly indifferent toward the behavior, and even increasingly to feel 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 preservice teachers were asked to indicate how appropriate they believed certain behaviors were. Only 1.4 percent thought that either changing answers on a student's answer sheet or giving hints or clues during testing were appropriate, and only 2.7 percent agreed that allowing more time than allotted for a test was acceptable. However, 8.1 percent thought that practicing on actual test items was okay, 23.4 percent believed rephrasing or rewording questions to be acceptable, and 37.6 percent judged practice on an alternate test form to be appropriate (Kher-Durlabhji & Lacina-Gifford, 1992).

The beliefs of preservice teachers appear to translate into actual practices when they enter the classroom. A large sample of third-, fifth-, and sixth-grade teachers in two school districts was asked to describe the extent to which they believed teachers in their schools practiced specific cheating behaviors. On the positive side, a majority of respondents said all but one of the behaviors listed occurred rarely or never (see Table 1). Equally noticeable, however, is that a wide range of behaviors was reported as occurring "frequently" or "often" by, in some cases, 15 percent or more of respondents. A second observation that leaps from Table 1 is the remarkable frequency with which teachers report that they have "no idea how often this occurs" (Shepard & Doughtery, 1991); this response suggests widespread unfamiliarity with other teachers' testing practices or lack of professional collaboration related to assessment.

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? Whereas 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. In addition, how might educator cheating affect students' attitudes toward tests or their motivation to excel? How might it affect their attitudes toward education, their trust or cynicism with respect to other institutions, or their propensity to cheat in other contexts?

Table 1. Teacher Beliefs About Inappropriate Test Administration Practices

Question: To what extent do you believe these are practiced by teachers in your school?	Percentage of Respondents				
Behavior	Never	Rarely	Often	Frequently	No Idea
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 studen'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 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 to those in the test	24.9	15.8	20.5	19.7	19.1

Recommended Strategies for Preventing Cheating

What can be done to deter cheating? Fortunately, many things. As a starting point, bringing the issue of cheating forward as a topic for discussion is likely to increase awareness of the problem by those who give and 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 identifying and addressing students' needs, we advance the perspective that obtaining accurate test results is more beneficial to all concerned than is cheating (Cizek, 1999, chap. 11).

Numerous more pragmatic steps can also be taken. The following list should provide a start. Of the following, some are focused on test givers, others on test takers, and some apply to both.

Get the Word Out

It has been said that we more often stand in need of being reminded than we do of education. 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 cheating on tests are often accompanied by protestations from the guilty parties 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.

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 student to copy another student's answer to an essay question, case analysis, or other constructed-response format than it is to copy a filled-in bubble response or to obtain the key to a multiple-choice item. Similarly, it is more difficult for an educator to forge or coach a student's answer to an essay question or a science experiment than to alter a filled-in bubble response or provide the key to a multiple-choice item.

It must be recognized, however, that a decreasing reliance on selected-response formats requires tradeoffs in terms of efficiency and scoring costs. It should also be recognized that the use of alternative formats will not completely solve the problem of cheating, for they can also be corrupted. (For an example of how the essay format can be corrupted on a state-mandated examination see Madaus, 1988).

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, 2000, 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, commentator 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).

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 student testing in the 1970s was made inevitable because of poor decision making—or at least the perception of poor decision making—and the resulting search for alternatives (see Cizek, 2001). 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 policymakers' pens.

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

Revise Test Disclosure Laws

States with so-called truth in testing laws or legislation requiring the release of secure test materials following their administration should reconsider the relative benefits of such laws. 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 of such laws to states have been staggering, because of the need to develop entirely new monitoring instruments one or more times each year.

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 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. Once delivered, materials should be maintained securely by a named person responsible for their security. After test administration, similar security procedures should be followed by those responsible for collecting, organizing, and shipping the materials.

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 proctors' vigilance or pursuit of instances of cheating. For all testing contexts, proper training must include instruction on methods examinees use to cheat, as well as 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.

Use Available Statistical Tools

Finally, recall that statistical detection methods should not be used to screen 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 Bellezza and Bellezza (1989) showed a reduction from approximately 5 percent to 1 percent in the incidence 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 of this potential use.

Enforce 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. Many tests are administered behind closed classroom doors 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 for cheating, including permanent disqualification from teaching within a state and more coordinated sharing of information regarding educators who have had their licenses revoked; and delegation of responsibility for investigating incidents of cheating to an independent authority.

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.

Summary and Conclusions

Overall, the evidence is in regarding the problem of cheating on tests: Cheating is occurring with increasing frequency. It is fair to conclude that the problem will not disappear. Therefore, it must be addressed 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 ways test administration staff may aid examinees in cheating. Additionally, those responsible for testing programs and those who oversee or give tests 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.

Note

1. A response to the Stancik report commissioned by the New York City teachers union called into question his methods and whether some of the accusations of educator cheating were based on credible evidence. The original report and subsequent response in this case highlight the serious nature of cheating allegations, illustrate the ambiguities surrounding the appropriateness of some practices, and recall the need to ensure that adequate guidelines and training regarding cheating are in place.

References

AERA, APA, & NCME [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,* pp. 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.

Bellezza, F. S., & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology, 16*(3), 151–155.

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 call 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 Associates.

Cole, N. (1998, November 9). Teen cheating hurts all. *USA Today,* 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), 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).

Loftus, E. (1979). *Eyewitness testimony.* Cambridge, MA: Harvard University Press.

Luecht, R. M. (1998). Testing and measurement issues: Automated test assembly in the era of computerized testing. *CLEAR Exam Review, 9*(2), 19–22.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum: Eighty-seventh yearbook of the National Society for the Study of Education* (pp. 83–121). Chicago: University of Chicago Press.

Mehrens, W. A., & Cizek, G. J. (2001). Standard setting and the public good. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 477–485). Mahwah, NJ: Lawrence Erlbaum Associates.

Merx, K. (2000, August 11). Cop test altered in Dearborn. *Detroit News,* A-1.

Messick, S. A. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.

National Association of Test Directors. (n.d.). *Appendix C: Testing code of ethics for North Carolina testing personnel, teachers, and school administrators.* Retrieved online from www.natd.org/Apndx_c.htm

Neill, M. (2000). Should high-stakes tests be abolished in order to reduce cheating? *Congressional Quarterly, 10*(32), 761.

Payne, P. (2000, August 18). Officials say 52 teachers paid $1,000 to pass competency tests. *[Schenectady, NY] Daily Gazette,* 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.

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

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

Shepard, L. A., & Doughtery, 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,* A-7.

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,* B-2.

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

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[ ] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[X] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").