Boston College

# Errors in Standardized Tests: A Systemic Problem

*National Board on Educational Testing and Public Policy*

Kathleen Rhoades & George Madaus
Lynch School of Education
Boston College
May 2003

NBETPP

# ACKNOWLEDGMENTS

We wish to thank the Ford Foundation for their encouragement and generous support of this project. We would also like to thank Walt Haney and Monty Neil for contributing reports on testing errors and related material. Finally, we are grateful to Anne Wheelock for the many testng errors that she brought to our attention over the past three years, and for the editorial comments she offered that undoubtedly improved our report. Any errors contained in this report are the sole responsibility of the authors.

# CONTENTS

# INTRODUCTION

## The Role of Testing in Society: The Imperfect Gatekeeper

Testing strongly affects our life and work. Educational testing results can open or close doors of opportunity from kindergarten through school and beyond, into one's job as a firefighter, sales clerk, or lawyer. Decisions based on state testing programs in elementary school can influence the type of secondary education one receives, and decisions at the high school level can affect one's path after graduation. All of these decisions are based on the quantification of performance – on numbers – and this bestows on them the appearance of fairness, impartiality, authority and precision.[1]

The heavy reliance of reformers on these numbers is a facet of what Thomas Aquinas calls a "cultivated ignorance," *ignorantia affectata*. Garry Wills (2000), writing in a different context, calls this ignorance "so useful that one protects it, keeps it from the light, in order to continue using it… this kind of ignorance [is] not exculpatory but inculpatory… a willed ignorance" (p. 9). Many proponents of high-stakes testing take a technological view: they choose to ignore the cumulative effects of test-based decisions, and view test takers as objects (Barbour, 1993; Foucault, 1979). Moreover, they ignore the fallibility of testing. Like any measurement tool that produces a number – whether a simple blood pressure reading or complex laboratory test – tests contain error. The widespread belief in their precision does not admit this inherent fallibility.

Like any measurement tool that produces a number – whether a simple blood pressure reading or complex laboratory test – tests contain error.

Two major types of error– random measurement error and non-random human error - are associated with tests. The first, random measurement error, is well documented and not treated in this paper. Briefly, it may be defined as "the consistency with which the [test] results place students in the same relative position if the test is given repeatedly" (Bloom et al., 1981, p. 76) so that tests with less measurement error produce more stable results than those with more.

This monograph is concerned with human errors, which differ from random measurement error in many ways. Human errors do not occur randomly; their presence is not known. These errors are of greater concern than random errors because they are capricious and bring with them unseen consequences. In contrast, measurement error is common to every test, and thus is expected; the amount of error is habitually calculated and disclosed, and therefore can be taken into account when interpreting test scores.

We will first examine human errors in systems in general, and in the education system in particular. We will then document active errors and latent errors in educational testing.

# Part 1: Errors in Systems, Errors in Testing

## Two Types of Human Error

An Institute of Medicine (IOM) report entitled To Err is Human (2000) divides human error into active and latent. Active error is most familiar: error committed by specific individuals, often easy to identify, and the first to be noticed. Latent error, by contrast, derives from misguided executive decisions (so-called bad management decisions). Latent errors "pose the greatest threat to safety in a complex system because they are often unrecognized and have the capacity to result in multiple types of active errors" (IOM, 2000, p. 55).

Latent errors "pose the greatest threat to safety in a complex system because they are often unrecognized and have the capacity to result in multiple types of active errors" (IOM, 2000, p. 55).

A recent example of latent error from England shows how latent error may contribute to active error. An audit committee there found that some 1,200 people died in public hospitals in 2001 because of mistakes in prescribing and administering medicine. This human error was brought about in part by latent error – by understaffing in hospitals and the increasing complexity of modern drug therapy –"which has created a culture where mistakes unfortunately do happen" (Lyall, 2001, p 4). A technological fix has been proposed; including computerized patient records and prescription systems, a standard national system for coding medicines, and the use of bar codes in the prescription system (Lyall, 2001).

In the K-16 educational testing system, active errors are often made by the testing contractor – for example, when a test item is scored incorrectly, a score conversion table is misread, or a computer programming error is made. These errors make news and therefore are familiar to the public.

Latent errors may stem from poorly conceived legislation or policy mandates, or from a faulty decision made at a department of education. For example, latent error in testing resides in the following:

- *Legislation requires a single test be used to determine graduation – a requirement that goes against the advice of the test developer or published test standards, or is in conflict with the test design.*
- *A state department of education demands that test scores be reported faster than can be dependably accomplished by the contractor.*
- *A mandate requires that school test scores increase by x percent per year. This directive fails to take measurement error into account, and projects growth estimates that are not realistic.*
- *Policy makers use tests to track achievement trends without any external confirmation of validity, making it impossible to tell whether the trend is due to actual differences in achievement or to some other factor, such as changes in the test.*

When active errors emerge, latent error may be implicated, but here one must be cautious. Substantiation of latent error can only be made after a systemwide inquiry is conducted. Such studies are rare in the United States.[2]

## Minimizing Human Error: A Systems Approach

To reduce human error, many technological systems seek to increase reliability by identifying problems in production or service delivery. This process, which involves the documentation of systemic weaknesses, resembles standard auditing procedures, but is continuous.

Medical and airline industries offer two examples of attempts at systemic quality control. In the medical arena, the IOM report (2000) defines human error as a systemwide, not an individual problem. The IOM recommends mandatory reporting of errors that result in injury or death, voluntary reporting of "near misses," and legislation to protect the confidentiality of individual reporters. This agency further recommends creation of a regulatory board to initiate research into errors within the system of medical care so as to reduce the nearly 100,000 medical errors reported to result in injury or death every year.

Likewise, the airline industry follows similar procedures. In a speech delivered to the National Press Club, David Lawrence (1999), MD and CEO of Kaiser Foundation Health Plan and Hospitals, spoke about the industry's development of, "a culture of safety that includes protecting those who provide information about 'near misses' and accidents from loss of job or legal action because of their willingness to report such information" (p. 3). Improvements in airline safety have been attributed to the adoption of these reporting procedures, as "from 1950 to 1990 commercial aviation fatalities declined from 1.18 to .27 per one million departures, an 80% reduction" (ibid., p. 2).

Leadership that communicates the value of reporting errors varies by industry. While the airline industry has encouraged full reporting, others (including the medical establishment and educational testing corporations) often classify their information as confidential and release only limited bits to the press. Moreover, unless safety is actively encouraged, workers have little incentive to report accidents. Consumer passivity compounds the problem. The IOM noted that "one reason consumers do not push harder for patient safety is that they assume accrediting and certifying organizations, as well as local and state regulators, do it for them" (Institute of Medicine, 1999, p. 3).

## Educational Testing: Difficulties with Detecting and Correcting Errors in a Closed System

In contrast to the airline industry where the reporting of mistakes is considered advantageous, and the consumer products sector where publications like *Consumer Reports* publicize errors, the testing industry is shrouded in secrecy (Mathews, 2000a). Its inner workings, in particular the arcane psychometrics of Item Response Theory (IRT) and standard setting, are outside the experience of most consumers. Thus the testing industry remains largely exempt from independent examinations. More than a decade ago The National Commission on Testing and Public Policy (1991, p. 21) pointed out:

> *Today those who take and use many tests have less consumer protection than those who buy a toy, a toaster, or a plane ticket. Rarely is an important test or its use subject to formal, systematic, independent professional scrutiny or audit. Civil servants who contract to have a test built, or who purchase commercial tests in education, have only the testing companies' assurances that their product is technically sound and appropriate for its stated purpose. Further, those who have no choice but to take a particular test – often having to pay to take it – have inadequate protection against either a faulty instrument or the misuse of a well-constructed one. Although the American Psychological Association, the American Educational Research Association, and the National Council for Measurement in Education have formulated professional standards for test development and use in education and employment, they lack any effective enforcement mechanism.*

Despite widespread use of testing in education and employment, there is no US agency that independently audits the processes and products of testing agencies. The lack of oversight makes errors difficult to detect.

Despite widespread use of testing in education and employment, there is no US agency (analogous to the Federal Trade Commission or the Federal Aviation Administration) that independently audits the processes and products of testing agencies. The lack of oversight makes errors difficult to detect. Individuals harmed by a flawed test may not even be aware of the harm. Although consumers who become aware of a problem with a test can contact the educational agency that commissioned it, or the testing company; it is likely that many problems go unnoticed.

Occasionally, consumers do notice problems with tests, and some lawsuits have ensued. One such court case was *Allen et al. v. The Alabama State Board of Education* (1999). Here, plaintiffs (Alabama teacher candidates) contended that the Alabama Initial Teacher Certification Testing Program (AITCTP) was biased against minority teacher candidates. Experts hired by the plaintiffs to determine the technical adequacy of the test needed a court order to gain access to the records of the testing contractor, National Evaluation Systems (NES). The plaintiffs' experts found items across various subject matter tests that did not meet

minimum standards for technical adequacy. In fact, they found miskeyed items[3] in tests that had been used over a six-year period (from 1981-1986). They showed that during one test administration, a miskeyed item resulted in at least six candidates failing who would have otherwise passed (Ludlow, 2001). Although this case was settled out of court, a verdict in a later trial rejected the use of the AITCTP on similar grounds. In this case, Judge Myron Thompson found many irregularities in the test development and pilot testing. One of these was the mathematical manipulation of the pass rate from one that was unacceptably low to one that was politically viable. Judge Thompson found that in making this change, the state "knew that the examinations were not measuring competency" (Richardson v. Lamar, 1989, p. 5). He further noted that the test development process lacked professionalism due to several serious errors the developer made in designing the 45 tests in 1981 and 1982. As a result, "errors at one step not only survived the next step, but also created new errors" (p. 7). In his decision, Judge Thompson wrote:

> *The [test contractor's] evidence reveals a cut-score methodology [4] so riddled with errors that it can only be characterized as capricious and arbitrary. There was no well-conceived, systematic process for establishing cut scores; nor can the test developer's decisions be characterized as the good faith exercise of professional judgment. The 1981 cut scores fall far outside the bounds of professional judgment (p. 7).*

Before this court case, the public and test takers knew nothing of these errors.

Judging a more recent NES teacher test has been equally difficult. The National Academy of Sciences commissioned the Committee on Assessment and Teacher Quality to evaluate the reliability and validity of the nation's teacher tests (Mitchell et al., 2001). The committee could obtain no usable data from NES.[5] Other researchers (Haney et al., 1999; Ludlow, 2001) have recently reported on their mostly unsuccessful struggles to access information from this company. This is highly troublesome because, as of today, NES is the nation's second largest producer of teacher licensure tests; in 2000, NES administered about 380,000 tests vs. 500,000 administered by the Educational Testing Service (Levinson, 2001). With regard to NES's practice of refusing to release technical data, the committee concluded:

> *The profession's standards for educational testing say that information sufficient to evaluate the appropriateness and technical adequacy of tests should be made available…. The committee considers the lack of sufficient technical information… to evaluate NES-developed tests to be problematic and a concern. It is also significant because NES-developed tests are administered to very large numbers of teacher candidates (Mitchell et al., 2001, p. 168).*

The committee recommended that tests not be used for making a decision of any import (such as decisions on teacher certification) if the test developer refuses to release critical data for determining test adequacy, as was the case here.

## Human Error in Testing: A Brief History

Testing errors are not new. In *The Testing Trap* (1981), Andrew Strenio documented several scoring errors that occurred between 1975 and 1980. For example, in 1978 a Medical College Admissions Test mistake resulted in artificially low scores for 90 percent of the test takers on one administration and probably caused some candidates to be disqualified. A 1977 error involved changes in the difficulty level of the Law School Admission Test (LSAT). Students who took the exam before October 1977 obtained lower scores overall, and therefore were less likely to be accepted into law school, than did those who took it after October. Strenio also documented an unusual error that occurred in 1975-1976. The Educational Testing Service (then publisher of the LSAT) "erroneously designated some law school applicants as 'unacknowledged repeaters,' i.e., persons who took the LSAT more than once" (Strenio, p. 14). Thus applicants who declared that this was their first time taking the test appeared to be lying, which "could hardly have improved their chances of admission" (p. 14). Strenio noted that without an oversight organization, the onus for protecting consumer rights usually falls to the testing contractors themselves:

> *In the end, we have to trust the companies to be as diligent in rooting out errors, as scrupulous in reporting their shortcomings, as skeptical of any unproven assumptions as an outside investigator would be. We have to trust the test companies to rise above their vested interests in protecting the reputation and sales of their tests by publicly calling into question their own performance whenever errors crop up. It is no slur upon the basic integrity and decency of the people working for the standardized testing industry to suggest that this is asking a lot (p. 15).*

Another famous testing error transpired from 1976 to 1980. During this time, a test-score calibration error resulted in the acceptance of more than 300,000 army recruits who normally would have been rejected because of low Armed Services Vocational Aptitude Battery (ASVAB) scores (National Commission on Testing and Public Policy, 1991). Sticht (1988) showed that, while the low-scoring recruits performed slightly less well than their higher-scoring counterparts, the difference was very small. Most of these low-scoring recruits performed well, many of them better than their higher-scoring peers. In fact, attrition for this period (a major performance indicator identified by the military as predictable by test scores) actually decreased slightly. This error may be considered a "naturally-occurring experiment" because it showed that those who "fail" tests might do notably better than predicted by test scores.

# Part 2: Documenting Active Errors in Testing

## Finding Active Error: Reporting Error to the Testing Contractors

In 1999, we began a systematic search for examples of testing errors. Over this three-year search, we found dozens of errors that were discovered by school officials, teachers, parents, and even students. Testing companies had not discovered most of them. Summaries of errors not discovered by testing contractors are contained in Appendix A. We describe below errors associated with three major publishers: CTB McGraw Hill, National Computer Systems, and Harcourt Brace.

### CTB McGraw Hill

In 1999, John Kline, director of planning, assessment and learning technologies for the Fort Wayne, Indiana, Community schools, reported a sharp drop in average percentile scores[6] on the TerraNova test to Indiana's Department of Education (DOE) (King, 1999). DOE officials asked CTB McGraw Hill, the contractor, to rerun the scores. McGraw Hill found that a programming error resulted in the use of the wrong table to convert reading comprehension raw scores to percentile scores. McGraw Hill adjusted the scores and sent out new results within two weeks (Viadero, 1999). Kline said at the time, "It's the largest glitch that we've ever seen, and I think we're going to find it's a bigger glitch than has been believed so far"[7] (quoted in Brunts, 1999, p. 1).

Two months later, William Sandler, a statistician working for the Tennessee DOE, questioned McGraw Hill about a score dip that affected two-thirds of Tennessee's TerraNova percentile scores. After weeks of wrangling, the company agreed to audit the data. Company officials were unable to diagnose the problem and simply adjusted Tennessee students' scores to correspond with Sandler's estimates (Steinberg & Henriques, 2001).

Over time, McGraw Hill determined that a programming error caused the percentile rankings on the TerraNova to be too low at the lower end of the scale and too high at the upper end. As a result, approximately a quarter of a million students in six states were given the wrong national percentile scores. In addition to Tennessee and Indiana, the error also corrupted scores in New York City, Wisconsin, Nevada, and South Carolina (Viadero & Blair, 1999). Students and staff in New York City and three Nevada schools were among the most seriously affected because officials used the scores to make high-stakes decisions. In the fall of 1999, officials from Nevada, New York City, and Indiana indicated that they were contemplating suing CTB McGraw Hill for costs incurred by the error (Zoll, 1999b; Bach, 1999b; Klampe, 1999).

Following discovery of the error on the TerraNova, CTB McGraw Hill posted a warning on their website: "No single test can ascertain whether all educational goals are being met" (Hartcollis, 1999d, p. 2).

To understand the impact of the error in New York City requires a review of decisions made in the months before the error was found. In early 1999 the TerraNova was used for the first time in New York City to measure achievement in reading and math (Hartocollis, 1999a). News of unexpectedly low scores hit the press that May. New York State Education Commissioner Richard Mills suggested that these new lower scores accurately reflected poor student performance and recommended that low-scoring students attend summer school that year (Hartocollis, 1999b). Driven by pressure from Mills and Mayor Rudolf Giuliani, Dr. Rudy Crew, then chancellor of New York City schools, chose the fifteenth percentile as the "cut point" below which children had to attend summer school or be retained in grade. Tens of thousands of New York City children scored below that point. Thirty-five thousand of them attended school that summer, and thousands of others were notified that they were to repeat a grade because they did not comply with the summer school attendance requirement (Hartocollis, 1999c).

On September 15, 1999, days after school started, McGraw Hill admitted that the same error that affected Tennessee scores had also incorrectly lowered New York City students' percentile rankings at the lower end of the scale. Thus, 8,668 children whose correct scores were above the cut-off had mistakenly been compelled to go to summer school.[8] In light of this new information, Dr. Crew announced that all the children affected by the error (many of whom had started school in a lower grade) would be allowed to advance to the next grade. Class lists were reorganized and children were rerouted to the proper classrooms in mid-September (Hartocollis, 1999c; Archibold, 1999). The irony in this story is that the TerraNova scores had actually *risen* substantially over those of the year before (Hartocollis, 1999d).

In Nevada, state officials used percentile rankings on the TerraNova to identify "failing" schools, which then received state funds for improvement. Recalculation of the scores showed that three schools were erroneously cited as "failing." They were allowed to keep the money due to the difficulty of returning funds already spent (Bach, 1999a). However, the negative publicity that accompanied inclusion on such a list could not be undone (Bach, 1999b). Following discovery of the error on the TerraNova, CTB McGraw Hill posted a warning on their website: "No single test can ascertain whether all educational goals are being met" (Hartocollis, 1999d, p. 2) – a strong caution against using test scores in isolation for making high-stakes decisions.

## National Computer Systems

In May of 2000, the daughter of a Minnesota lawyer learned that she had failed the math portion of Minnesota's Basic Standards Tests (BSTs), a test published by National Computer Systems (NCS). Her father contacted the Department of Children, Families and Learning (CFL), asking to see the exam. For two months CFL staffers rejected his request and "told him to have his daughter study harder for next year's exam" (Welsh, 2000, p. 1). Only when the parent threatened a lawsuit did CFL permit him to examine the test (Grow, 2000).

For two months CFL staffers rejected his request and "told him to have his daughter study harder for next year's exam" (Welsh, 2000, p. 1).

The father, along with employees from CFL, found a series of scoring errors on Form B of the math test administered in February 2000. The errors were later traced to an NCS employee who had incorrectly programmed the answer key (Carlson, 2000). As a result, math scores for 45,739 Minnesota students in grades 8-12 were wrong. Of these, 7,935 students originally told they failed the test actually passed (Children, Families, & Learning, 2000a). Another error involving a question with a design flaw was found on the April administration of the BSTs. NCS invalidated this item, but not before 59 students were erroneously told they had failed (Children, Families, & Learning, 2000a).

Since passing the BSTs was a requirement for graduation, more than 50 Minnesota students were wrongly denied a diploma in 2000. Of this number, six or seven were not allowed to attend their high school graduation ceremonies (Draper, 2000). The State Education Commissioner expressed deep regret over the incident, saying, "I can't imagine a more horrible mistake that NCS could have made. And I can't fathom anything that NCS could have done that would have caused more harm to students. I know that NCS agrees with me" (Children, Families, & Learning, 2000b, p. 1).

CFL tried to rectify the mistakes by providing a telephone consultation service to parents. By August 2000, concerned parents inundated the line with calls seeking clarification of the error and its impact. NCS agreed to submit to an outside audit under threat of losing the contract with Minnesota, and offered to pay each affected student $1,000 toward college tuition as well as all out-of-pocket costs, such as tutoring and mileage expenses, resulting from the error. In all, NCS paid Minnesota families $118,000 and CFL $75,000 for costs related to the error (Drew & Draper, 2001). While NCS took full responsibility for the errors, CFL reprimanded a staff member for failing to return the father's original e-mail message. Had it been answered promptly, the error might have been caught before students were barred from graduating (Drew, Smetanka, & Shah, 2000). State senators also criticized the Commissioner's office for providing inadequate oversight of the testing program (Grow, 2000).

In spite of the test company's attempts to remedy the situation, four parents sued NCS, claiming that the company "was aware of repeated scoring errors and quality control problems" (Corporate greed, 2002). In a court ruling in the summer of 2002, the judge denied the plaintiffs' request for punitive damages because he decided that NCS had not intentionally produced a faulty test (Welsh, 2002, p. 1). The judge reconsidered his ruling in the fall of 2002 in light of compelling evidence brought forth by the plaintiffs. He determined that the plaintiffs could seek punitive damages against NCS because they "produced prima facia evidence that [NCS] acted with deliberate disregard for the rights and safety of others" (Grant, 2002; Kurvers et al. v. NCS, Inc., 2002, p. 4). His reason for repealing the first verdict was:

> *[The] Plaintiff Class has presented volumes of evidence detailing the systematic failure of NCS' quality control systems. Although it appears that the mistake that led to the scoring error was simple, Plaintiffs have demonstrated that the error was preceded by years of quality control problems at NCS (p. 1).*

"[The] Plaintiff Class has presented volumes of evidence detailing the systematic failure of NCS' quality control systems" (Kurvers et al., v. NCS, Inc., 2002, p. 1).

He attributed the causes of NCS' problems to managerial error instigated by a profit-seeking ethic at NCS that prevailed over other consideration for the customer:

> *NCS continually short-staffed the relatively unprofitable Minnesota project while maintaining adequate staffing on more profitable projects like the one it had in Texas. This understaffing and underfinancing occurred during a time in NCS' history when management was attempting to increase profits (Kurvers et al. v. NCS, Inc., 2002, p. 2).*

Indeed, within days of acknowledging the error, Pearson, an international distributor of educational materials, purchased NCS. Following the $2.5 billion dollar acquisition, NCS stock rose dramatically and provided the CEO with millions of dollars in profit (Wieffering, 2000). Before the case went to trial, however, NCS settled with the plaintiffs for $7 million dollars, paying all of the students who missed graduation $16 thousand each (Scoring settlement, 2002).

## Harcourt Brace

Another error, this time by Harcourt Brace, occurred against the backdrop of California Proposition 227. Prop. 227 required California schools to educate children classified as "limited English-proficient" (LEP) in English-speaking classrooms. It also mandated other changes, including smaller class sizes and curriculum changes. Policy makers selected Harcourt's Stanford 9 Achievement Test (SAT-9) to gauge the achievement of students across the state, asserting that gains on this test would signal the success of 227 (Sahagun, 1999). Initial results reported at the end of June 1999 did show large gains for students in English immersion classes with scores of LEP students rising by as much as 20% in some schools (Mora, 1999). Proponents of English immersion quickly touted these gains as an indication of the Proposition's success (Brandon, 1999; Colvin & Smith, 1999). Within days, however, the state discovered that 257,000 newly English proficient students had been misclassified as LEP (Stephens, 1999). Once these students were correctly classified, test score gains of LEP students were substantially reduced (Sahagun, 1999). LEP students' scores had risen, "slightly in most grades, but still were far below the national average" (Chrismer, 1999, p. 1).[9] Further, because it was the second year the state had administered the SAT-9; small gains due to familiarity with the test had been anticipated (Orr, Butler, Bousquet, & Hakuta, 2000).

This error, like the McGraw Hill and NCS errors, was detected by consumers and not the test publisher. However, unlike McGraw Hill, Harcourt Brace reanalyzed the data quickly. Then Harcourt president Eugene Paslov noted, "[The error] might have been caught if the company had more than two days to analyze data from 4.3 million test forms… before Wednesday's deadline for posting results on the Internet" (quoted in Colvin & Groves, 1999, p. 2). The observation is an excellent example of a latent error evolving into human error – as a faulty policy decision resulted in insufficient time to process the tests and caused an active error by the contractor.

Moreover, officials from California's DOE reported very little state oversight of the testing program. Gerry Shelton, an official with California's DOE, had lobbied for an $800,000 office that would verify the accuracy of test results. Shelton cautioned, "The department doesn't have any responsibility… because the publisher is running the program. This specific problem could have been "prevented with an hour of time from one of my staff members (Colvin, 1999a, p.2)." An unnamed source from Harcourt Brace echoed Mr. Shelton: "There is no one in charge. This thing is out of control. You've got the California Board of Education, competing voices in the state Department of Education, the governor – and 1,100 separate contracts" (Asimov, 1999a, p. 1). A top official from Harcourt Brace warned, "We can't check the results for each of the 1,100 districts. It's invaluable for folks at the local level … to check to see if there's anything that looks suspicious" (quoted in Colvin, 1999b, p. A3).

Despite the scoring problems of 1999, California Governor Gray Davis signed a bill in October of that year awarding a single state contract to a testing company. The bill called for rewards or penalties for schools based on a list of criteria (test scores prominent among them) and opened the door for school takeovers by the state, retention in grade, and denial of diplomas on the basis of test scores (Asimov, 1999b). Further, the bill mandated that test results be published by June 30. These provisions incorporated latent error and so increase the probability of active errors in the future. Indeed, Harcourt officials attributed errors made in 2001 to the cramped time table for reporting results and indicated that in the future they would rather be penalized than send out results that went unchecked (see Appendix C, #13) (Groves & Smith, 2001).

"There is no one in charge. This thing is out of control. You've got the California Board of Education, competing voices in the state Department of Education, the governor – and 1,100 separate contracts" (Asimov, 1999a, p. 1).

## We've Made an Error…Errors Found by Testing Companies

In a number of instances, testing companies contacted clients to report an error. In one such case the Educational Testing Service (ETS) identified an error on a 1999 administration of the SAT that lowered the scores of 1,500 students by as much as one hundred points[10] (Sandham, 1998; Weiss, 1998). In a similar occurrence in the United Kingdom, the Qualifications and Curriculum Authority (QCA) quickly discovered a computer programming error that resulted in over 10,000 students being misclassified on the secondary school information technology exams (Students given wrong test results, 2001). These errors were detected in the course of standard auditing procedures. Companies that regularly audit results, like ETS, are more likely to detect and correct errors both before and after results are sent out to the public. Appendix B contains summaries of errors found and reported by testing contractors.

## One Question – the Difference between Success and Failure

Errors occur when test questions are ambiguous, poorly designed, or miskeyed. On a high-stakes exam, one poorly written question can determine a student's classification as "passing" or "failing," and reversal is difficult.

On a 1987 administration of the New York Bar Exam, Steven Shapiro challenged one ambiguous question, which was found to be faulty and thrown out. When the exam was rescored, twenty-six individuals learned that they had passed after being told that they had failed. Exam scores in 1985 and 1986 also had to be changed because of errors. In July 1985, 111 failing grades were reversed because of a scoring error on the exam (New York Bar Exam, 1988).

Removing a question from a test is far from a simple remedy. First, K-12 students are unlikely to remember individual questions, particularly on a long exam, so it is often up to educators to challenge faulty questions for them. Such a challenge came from John Anderson, a high school teacher from Bell Buckle, Tennessee, who in 1988 found an ambiguous English question on the Tennessee High School Proficiency Test (THSPT). At that time, public school students had to pass the THSPT to obtain a high school diploma. Mr. Anderson noted that a few students had failed the test by one question – the same one – and that they had chosen another viable (but "wrong" according to the key) answer for that item. State officials turned down Mr. Anderson's request to review the question, stating that 98.6% of the state's students chose the keyed response, which, therefore, must be correct. Mr. Anderson then sought the opinions of experts in English (including Claire Cook, then Editorial Supervisor of MLA Publications) and in psychometrics. Both agreed with Mr. Anderson's objection and relayed their positions to Tennessee's Department of Proficiency Testing. The director of the department at the time indicated that grammarians consulted within the department stood by the keyed response. She also continued to maintain that the correct-response rate of 98% essentially nullified Mr. Anderson's complaint. A reply from Dr. George Madaus, however, called that assumption "erroneous" on the grounds that item difficulty levels do not necessarily reflect item accuracy. He further commented:

> *What we seem to have is two groups of experts, the State's and Mr. Anderson's, and they are in disagreement about the correct answers. I am in no position to decide which group is correct, but given the disagreement, students should receive the benefit of the doubt (G. Madaus, personal communication, December 8, 1987).*

The incident concluded with a ruling against Mr. Anderson's claim. Student failures remained fixed (Dickie, 1987; J. Anderson, personal communication, November 28, 1987).

More recent examples of how one question can determine passing or failing have been found in Arizona, Virginia, Massachusetts, and Nevada (Appendix A, # 29 , 47, and 52, and Appendix B, # 24, respectively). The Massachusetts error, identified by a student who had already failed the Massachusetts Comprehensive Assessment System (MCAS), involved a multiple choice question that had more than one correct answer. Since the item was part of a retest taken by juniors and seniors who had not yet passed the high school exit exam, 449 of those who chose the alternate answer and had previously failed the test were found to be

"competent" and worthy of a high school diploma – so in this case, the one question served as an arbiter of who would and would not graduate (Massachusetts Department of Education, 2002b). Ironically, the student who discovered the second correct answer still had not obtained enough points to pass the test (Kurtz & Vaishnav, 2002). In Virginia and Nevada, errors were a result of test equating – a statistical process used to make the scores of a test comparable from one year to the next. Through that process, testing contractors determine the points needed for passing the current year's test that is equivalent to the passing score on that test the year before. In both cases, the passing score was set one point (or approximately the equivalent of one question) too high (Akin, 2002; Ritter, 2002)[11] And in Arizona, after state educators identified a miskeyed question, the scores of 12,000 high school sophomores increased and 142 students who failed the test, passed (Pearce, 2000a).

Another type of error is the faulty test question, found by test-takers or proctors during a test administration, and later removed before scoring. For example, during the 2001 administration of the MCAS, two tenth-graders found one math multiple choice item where all of the answers provided were correct (Lindsay, 2001). One of the students reported that he worked more than five minutes on an item that should have taken one or two minutes at the most. Test officials often claim that removal corrects the problem since the removed item does not affect the scores. This does not correct the disruption experienced by knowledgeable test-takers who try to find a solution to a faulty question. Indeed, Michael Russell, a professor at Boston College, suggests an analogy that demonstrates the residual impact of faulty items, even if they are removed from consideration during scoring. During the 2000 summer Olympics, Svetlana Khorkina, international favorite to win a gold medal in the women's gymnastics competition, failed to win the coveted prize. Khorkina ranked first after scoring a 9.812 on the floor exercises. She then moved to the vault where she did a most uncharacteristic thing – she landed on her hands and knees (Harasta, 2000). After a string of similar mishaps, event officials rechecked the vault's height. It was set at 120 instead of 125 centimeters – a difference of a little less than two inches. The vault was quickly set correctly and the affected gymnasts were allowed to repeat their vaults, but the damage was done – Khorkina was unable to regain her momentum or her confidence, and declined another attempt on the apparatus. She left the event in 10th place, far behind her initial standing, and ended the competition in 11th place after a fall on the uneven parallel bars (Measuring mix-up, 2000).[12] Russell suggests that test takers, confronted with a question that doesn't make sense, may suffer a similar loss of confidence. And some students may spend an inordinate amount of time puzzling over the flawed, and eventually discarded, item and so have less time for the remaining questions (M. Russell, personal communication, July 11, 2001).

## Something Wrong with the Rankings

As of 2002, more than two dozen states ranked school or district performance by test scores, and in twenty of these states sanctions could be levied against low-scoring schools (Meyer et al., 2002). For example, the DOE in Virginia uses scores from the Standards of Learning tests (SOLs) to assign schools to ranks, which then determine school accreditation. Schools must earn accreditation by the year 2007 (Seymour, 2001).

As in other states, Virginia's ranking system is vulnerable to test score error. In October 2000, soon after SOLs results were released, administrators from the Virginia Beach school department challenged the ratings of several elementary schools, which were lower than projected. Upon investigating the complaint, the DOE admitted that an omission[13] had produced incorrect rankings for possibly dozens of schools. After the low ratings at Virginia Beach were upgraded, a spokesperson for the district said, "That's three schools that received bad publicity, and they didn't deserve it" (quoted in Warchol & Bowers, 2000, p. 1).

The Virginia example is not an isolated incident. As test scores are used increasingly to rank schools and award teachers' bonuses, the likelihood of making mistakes also increases. Recent errors in school rankings are summarized in Appendix C.

## Gray Areas – When Test Results Don't Add Up

The Stanford 9 Achievement Test (SAT-9) is a widely used norm-referenced test.[14] From 1998-2000, a perplexing scoring trend has been exhibited on Form T[15] of this reading test, but not on its parallel form, Form S. During this time, education officials from six states (Alabama, Arizona, California, Delaware, Florida, and South Dakota) reported a distinct and inexplicable drop in student scores at grades nine and ten (Hoff, 2000; Schrag, 2000). An unexplained dip in scores would present a problem in any event; and stakes attached to scores magnify the consequences. For example, in California, teacher bonus pay is based on SAT-9 results (Policy & Evaluation Division, 2000). In Florida, schools must report SAT-9 scores to parents; and the Florida DOE makes the scores available to the public via the World Wide Web (Florida Department of Education, 2001).

Officials in both Florida and California have repeatedly questioned the results from Form T (Nguyen, 1998; Smith, 1998; Groves, 1999; Hegarty, 2000a & 2000b). In Florida, for example, average reading scores in 2000 plummeted from the 54th percentile in grade eight to the 38th and 33rd percentiles in grades nine and ten, respectively (Hegarty, 2000b). As a result, the Education Commissioner delayed release of the scores by three months while the state's Inspector General's Office (IGO) investigated. In September, the IGO issued a report ruling out errors in scoring or programming. It could not pinpoint the cause of the lower Form T scores (Hegarty, 2000b; Hirschman, 2000). Results from an alternate form – Form S – showed no such anomaly (Nguyen, 1998; Smith, 1998; Groves, 1999; Hegarty, 2000a & 2000b).

Similar anomalous percentile rankings on Form T in California have caused assessment officials to question the high school reading results. Among the possible explanations proposed were: (a) the students didn't take the test seriously, (b) changes in state reading curricula and instruction led to lower achievement, (c) students didn't read enough at home, and (d) the norm group for Form T was composed of students with unusually high reading achievement. However, only (d) would explain why Form T exhibits the abnormality and Form S does not. Officials were reluctant to replace Form T with Form S because all of the state's

baseline data were drawn from Form T. Although California officials adopted a number of strategies to remedy the poor performance in reading that included more structured reading programs as well as remedial reading classes (Nguyen, 1998; Smith, 1998; Groves, 1999), the score dip remained.

Officials at Harcourt Brace, publisher of the SAT-9, have consistently denied that their Form T is flawed. Thomas Brooks of Harcourt Brace said, "We've done everything we can. As far as we've been able to determine the procedures have been followed correctly" (quoted in Hoff, 2000, p. 1). In fact, low scores may reflect an error in the norming process. If the norm group was not representative of the student population, and especially if it consisted of a group of high achievers, then the students taking the test would tend to score low in comparison.

This problem of score validity has arguably been most serious in California where SAT-9 scores heavily determine teacher bonuses and school rankings. Yet, because the SAT-9 test scores were so firmly entrenched in this accountability system, the test that produced what seemed to be spurious scores could not be easily replaced. Peter Schrag noted:

> *If this were a case of faulty tires, there would be talk of a recall. But since the states that use the SAT-9 have now established it as a base with which to compare future performance, each has a political and financial investment, ironically maybe even an educational one, in staying even with a flawed test (Schrag, 2000, p.3).*

Other state departments of education have struggled with unusual test score patterns associated with cut scores set for their assessment programs.[16] Over the first three administrations of the Massachusetts Comprehensive Assessment System (MCAS) fourth-grade English/Language Arts (ELA) scores were incongruously low with 80 percent of Massachusetts' fourth graders scoring in the lowest two of four categories in reading each year. Several factors pointed to the possibility that cut scores had been set in error. First, the pattern of low scores was not repeated on the fourth-grade math and science tests (Griffin, 2000; Massachusetts Department of Education, 2000). Second, only 38% of Massachusetts' eighth-grade students scored within the lowest two performance levels on the eighth-grade English/Language Arts test (Massachusetts Department of Education, 2000). Third, Massachusetts had one of the lowest percentages nationally of students scoring in the bottom two performance categories on the National Assessment of Educational Progress (National Assessment of Educational Progress, 1998). These discrepancies prompted state officials to reconsider the performance levels established in 1998 (Driscoll, 2001) and eventually adjust them. Adjusted fourth-grade 2000 scores reported with the 2002 test results showed that instead of 67% of students scoring in the "needs improvement" category and 13% scoring in the "failing" category (as was reported in 2000), 35% now scored as "needs improvement" and 16% scored as "failing" – a significant shift in scores (Massachusetts Department of Education, 2002a). See Appendix D for a fuller account of these unexplained test score patterns.

# Part 3: Documenting Latent Error in Testing

## Tracking Trends: A Process Susceptible to Latent Error

As of 2001, 40 states and numerous large cities were using trend data to hold schools and districts accountable for educational achievement (Kane & Staiger, 2001). Trend is usually determined by simply aggregating and averaging the test scores of individual students and comparing them from year to year. Yet this process, which appears to be simple, is actually very complex.

NCES found that six test items (more than found on most state writing assessment programs) were too few to reliably link the scores.

Tracking trends involves a decision-making process that is particularly susceptible to latent error. Very often, procedures designed to track trends can fail to account for other factors or sources of instability that affect students' test scores. These factors may include: changes in the student populations being tested, changes in the test (such as the addition of new test questions and the deletion of old ones), and changes in the conditions under which examinees sit for exams. In order to generate accurate trend data, procedures used to track trend should incorporate methods designed to ameliorate the effects of sources of instability.

In a recent report, Kane and Staiger (2001) used five years of data from North Carolina's accountability program to determine trends for grades three through five. They found that two sources of variability interfered with tracking changes in student achievement. First, the small number of students tested per school (about 60 or less), coupled with changes in student composition from year to year, did as much to shift school means as did real changes in student achievement. Second, random variation – events taking place during test administrations, such as disruptions, illness, or the relationship between student and teacher – were found to affect scores. Kane and Staiger recommended that accountability programs filter out as much of this variation as possible.[17]

Even when random variability is filtered out, tracking achievement trends is tricky, even in the most carefully designed assessment programs, as in the case of the National Assessment of Educational Progress (NAEP). Many researchers use NAEP, also known as the Nation's Report Card, to assess the validity of state and local test score trends. That is because since its inception in 1969, one of NAEP's goals has been to collect long-term trend data on US achievement using two procedures. Each NAEP assessment contains a core of questions that are re-used across test administrations; and a statistical process known as 'bridging' is used to link the scores statistically from year to year.

In 2000, however, the National Center for Education Statistics (NCES) surprised many when it removed the NAEP long-term writing results from its web site, a decision made after NCES detected several problems with measuring change in writing achievement (National Center for Education Statistics, 2000). The first of these was an insufficient number of test items. In order for tests to validly measure writing ability, students must write; but because writing long essays is time-consuming, the number of questions that can be asked is limited. NCES found that six test items (more than found on most state writing assessment programs) were too few to reliably link the scores and concluded that the trend data for writing could not be trusted.

The second problem, called scorer drift, involves scoring inaccuracies on open response items that worsen over time. While it is essential to avoid scorer drift, it is difficult to do so. Gary Phillips, Acting Commissioner of NCES, opined, "While we can develop standards for scoring students' answers, train scorers, establish reliability and monitor the process, it is unrealistic to think that scorer drift will ever completely be eliminated" (National Center for Education Statistics, 2000, p. 1).

A third problem with measuring change in writing achievement lies in the use of writing prompts. Using the same prompts in different test administrations allows the NAEP assessment to measure change, but the conditions under which the prompts are received are constantly changing. For example, a prompt that asks students to think of a significant event that affected their lives may be answered in particular ways by students at different times, depending upon what is happening in the world around them. These forces introduce instability into students' responses, which greatly complicates the measurement of change.

While the problem of tracking trend data using open-ended items has yet to be solved, another earlier problem with a NAEP administration shows that even subtle changes made in tests between administrations can obfuscate real changes in achievement (Beaton et al., 1990). A NAEP report on reading achievement, set for release in 1986, was not issued until 1988 due to an unanticipated problem with the trend data. Despite the meticulous processes used to bridge NAEP results of different test administrations, small changes in the 1986 reading test created dramatic – but invalid – differences in student results between that test and earlier ones. Initial results from the 1986 reading test showed that aggregate scores for 9- and 17-year-olds had shifted conspicuously downward from 1984, while scores for 13-year-olds were in line with past trends. The downward trend for 9- and 17- year olds was larger than all of the other trend changes between any two- to five-year period in the history of NAEP's trend studies, and remained even when the data were broken out by gender, race/ethnicity, parents' education level, and region.

Despite the meticulous processes used to bridge NAEP results of different test administrations, small changes in the 1986 reading test created dramatic – but invalid – differences in student results between that test and earlier ones.

As a result, researchers conducted a series of explorations to determine the reason for the downturn in the 1986 scores. Hypotheses that were entertained then discarded included: sampling problems, problems with specific test items, and computational errors. Each in succession was ruled out, with the exception of the possible effect of a few minor changes in the 1986 reading test – changes thought to be too insignificant to appreciably alter results. The changes included the following:

- *The number of items per test booklet was increased slightly along with the total time allotted to administer each block of items.*[18]
- *A tape recording to pace test administration (generally used for all subject matter tests) was not used for the reading portion of the test. As a result, many students had less time to complete the items that appeared later in the reading block.*
- *The composition of the test booklets was changed: in some test booklets, reading items were combined with math and science items, while formerly they had appeared in booklets with writing items only.*
- *Students were instructed to fill in a bubble to mark the correct answer, rather than circle the answer as they had done in the past.*

Statisticians tested the effects of these changes during the 1988 NAEP exams in an experiment involving the administration of the test to two additional samples of students. One group of students was given the 1984 NAEP tests, while a second group used the 1986 test booklets and administration procedures. Since both groups were randomly sampled from the same population, score differences could be attributed to test design. The experiment showed that small design changes did little to affect the accuracy of measurement for individuals; however, when scores were aggregated to measure group change, small errors were compounded, creating large errors in the scores used to track progress over time.

A study on the effects of small changes in science performance assessments tells a similar story (Stetcher et al., 2000). Researchers sought to determine how scores on science test items varied when changes were made in (1) item content (different tasks sampled from physical/chemical sciences); (2) item format (paper-and-pencil vs. hands-on tasks); and (3) the level of inquiry (unguided vs. guided tasks). The correlations between similar items (in content, format, or level of inquiry) were not significantly higher than the correlations between dissimilar items. The investigators noted, "Whatever the reason, the lack of differentiation in scores due to format, content, or level of inquiry in our research raises important questions about what open-ended tasks in science truly measure" (p. 154). In addition, the authors found that the use of different teams of people to design test items introduced a significant a source of variability. They cautioned:

> *This study also speaks more generally to the difficulty of developing complex performance assessments. It is worth noting that the individuals involved in this project had many years of experience developing tests in a variety of subjects and formats. Furthermore, we used a thoughtful, unhurried approach, including pilot tests and revisions. Nevertheless, the results of our efforts did not meet our expectations. We would recommend that test developers increase the amount of pilot testing that is done on performance assessments and that pilot tests be designed to investigate whether performance on one task is affected by the completion of other tasks (i.e., order effects) (p. 154 - 155).*

In 2002, problems with trend data emerged in Nevada, Georgia, and Virginia. In each of these states, Harcourt Educational Measurement (formerly Harcourt Brace) officials announced that errors were made in test equating so that in each state the test scores were seriously affected. In Nevada and Virginia, use of a new computer program to equate the 2002 with the 2001 scores resulted in the cut score on the 2002 test being set one point too high. In Nevada, this meant that 31,000 high school graduation scores were incorrectly reported and 736 students were told they had failed when they had passed (Hendrie & Hurst, 2002; State despairs of getting, 2002). In Virginia, thousands of students were erroneously told they had failed or were given test scores that were too low (King & White, 2002). And in Georgia the results were so flawed, and so late that they were deemed unusable. Making

matters worse, officials from Harcourt could not determine the nature of the problem in Georgia since these results emanated from standard equating procedures (Donsky, 2002). These problems in equating convey a degree of uncertainty regarding the accuracy of trend scores obtained under such methods.

Harcourt was also connected to two recent examples of dramatic score increases – the 2000 – 2002 Standards of Learning (SOLs) tests in Virginia and the 2001 tenth-grade MCAS tests in Massachusetts. In Virginia, where the tests are used to accredit schools, the percentage of students passing the exit exams increased sharply each year, from only 6.5% of the schools meeting accreditation standards in 1999 to 23% in 2000, 41% in 2001, and 66% of all schools passing in 2002[19] (Benning & Mathews, 2000; Helderman & Keating, 2002). These increases coincided with new regulations that required schools to have high pass rates on the SOLs in order to maintain accreditation in the year 2007 (Seymour, 2001) and graduation requirements that linked graduation to passing a series of exit exams (Helderman & Keating, 2002). Similarly, in Massachusetts, the scores of the first class of tenth-graders who had to pass the MCAS to graduate rose substantially among all categories of students (Hayward, 2001b; Massachusetts Department of Education, 2001). Passing rates in math jumped from a three-year range of 55-62% to 82% in 2001; in English the increase was less dramatic – after a steady decline in pass rates for three years (1998: 81%, 1999: 75%, 2000: 73%) the rate rose to 88% (Greenberger, 2001; Gehring, 2001). The large tenth-grade increase in 2001 pass rates was maintained in 2002 with another slight increase in English rates and a minor decrease in math rates (Massachusetts Department of Education, 2002a).

Harcourt could not determine the nature of the problem in Georgia since these results emanated from standard equating procedures.

Some attributed the large increases in both states to the urgency brought on by the new requirements for both students and teachers, that both worked harder because so much was riding on the test (Gehring, 2001; Greenberger & Dedman, 2001; Mathews, 2000b). Others pointed to changes in each testing program that contributed to increases (O'Shea & Tantraphol, 2001; Hayward, 2001b; Greenberger, 2001; Helderman & Keating, 2002). Indeed, changes in scaling drove the Massachusetts Commissioner of Education to warn the media that scores would go up slightly, so that they would not overinterpret small gains (Collins, 2001).[20] The changes in pass rates in both Massachusetts and Virginia could not be characterized as small, however, leaving open the question whether the sharp increases reflected real gains in achievement or resulted from test modifications (changes in the items, the equating procedures, and other statistical procedures, including scaling) or reflected a combination of factors including the attrition of low-scoring students from the test-taking pool.[21] Casting more doubt on large score increases are the small, statistically insignificant increases observed in the NAEP long-term trend data from 1996 and 1999 in reading and math (increases ranging from 0-2 points in a 0-500 scale) (Campbell et al., 2000).[22]

In the U.K., maths test examiner Jeffrey Robinson was skeptical of large increases observed over the past decade on GCSE maths exam scores (Smithers, 2001). He conducted a fourteen-year examination of the correspondence between raw scores and test grades and reported, "Standards in maths had dropped by two grades in the past ten years" (p. 1). Upon presenting statistical evidence to back up his claim, Robinson asked for an outside examination. Before investigating, a UK test examiner tried to explain the score shifts in this way:

> *The plain facts are that if we are to challenge high, intermediate and lower ability pupils properly, the questions and the marks set for those questions will change over time….. When questions are made more difficult, obviously a lower mark will represent the same ability level as a higher mark on easier questions. That is what happened at GCSE, put in an over-simplified way (p. 1).*

He thus connected shifts in scores to changes in the test. A subsequent examination by the Qualifications and Curriculum Authority failed to corroborate Robinson's findings (Qualifications and Curriculum Authority, 2001). Approximately three months after making his accusation, Robinson was fired by the Oxford Cambridge and RSA Examinations board for "breach of contract" (Examiner sacked, 2001, p. 2). A former chief inspector of schools, Chris Woodhead, defended Robinson by noting that others in the field agree with him: "Jeffrey Robinson is saying publicly what a number of chief examiners have said privately over the years" (Waterfield, 2001, p.1).

All of these events and investigations counsel caution in the interpretation of changes in scores from year to year, as these could very well be the result of changes in the tests and not actual changes in achievement. What we learned from the NAEP and Stetcher et al. experiments is that even very small modifications in tests can yield dramatic changes in results and that one cannot rule out extraneous causes without an outside audit of the results. Even given the careful design and construction of items, extraneous variability crops up, making it difficult to compare the scores of students on apparently similar open-ended items from one year to the next. NAEP was designed to measure long-term trends, and individual scores are not reported. The NAEP instruments, therefore, can re-use items more frequently than can state assessments. Most of the latter (TAAS, MCAS, FCAT, etc.,) use a test once or twice and then release the majority of items to the public.[23] Released items are then replaced, which can alter the test in significant but unintended ways and make the detection of change an uncertain endeavor. Further, the use of statistical methods to link scores on tests from year to year, while necessary, may, in themselves, introduce additional variability in the measurement of trends. Therefore, because of the many pitfalls inherent in measuring change, such endeavors must be approached with the greatest caution. In determining the nature of changes in trend scores, therefore, one must always be on the lookout for "plausible rival hypotheses"[24] as there often are a great many of these hypotheses to entertain.

## An Example of Latent Error: The Scottish Qualifications Authority

One of the most thoroughly investigated cases of latent error in educational testing was examined in 2000 by the Scottish Qualifications Authority (SQA). It illuminates the nature of human error and the role that education systems can play in creating an atmosphere in which errors are likely to occur.

The SQA was formed in 1997 to combine the former General Education and Vocational Education Councils. Among its responsibilities was the testing of candidates for Scotland's universities. In Scotland, students must pass a series of exams to be admitted to university. Unlike other national testing programs that make students who fail exams ineligible for admission to post-secondary education, Scotland's testing program is both more ambitious and more forgiving. If students fail they may take the exams in subsequent years, even as adults, to earn a space at a university (Scottish Qualifications Authority, 1997).

The SQA also directed the rest of the country's K-12 assessment system. This program stretched available resources, which had been level-funded for years. Hundreds of markers[25] were needed to score the increasing number of exams because more students were sitting for exams each year; and the pay was low. A tight exam schedule meant that only a few months separated the marking of exams from the publishing of results. Schools, meanwhile, were expected not only to provide markers, but also to supply the SQA database with predicted scores[26] for every student. When marks came out in mid-summer, students only had about a month to appeal their grades before they lost a coveted space at a university for that year. For this whole process to work, it had to operate within a very slim margin of error (Henderson & Munro, 2000; Macdonald, 2000a; MacBride, 2000).

In the summer of 2000, exam scores arrived late, and inaccuracies were reported. By mid-August, over 5,000 potential university students had received incomplete or inaccurate results (Macdonald, 2000b). University placements (admittance decisions) were down by almost 9%, prompting Scottish Education Minister Sam Galbraith to promise that, "No Scottish student would miss out on a university place," although he conceded that students might not be granted their first choice (quoted in Mackay, 2000, p. 3).

By the end of the summer more than 4,600 students had appealed their grades because they were out of line with students' and teachers' predictions (Mackenzie, 2000). James Dalziel, head teacher at Eastbank Academy, Glasgow, described the situation:

> *I know that some people are saying that head teachers were being a bit alarmist about this, but we've all been around for quite a long time and never before have we seen such glaring inconsistencies or … such variance between the estimated grade of the teachers and our pupils' results (quoted in Mackay, 2000, p. 4).*

Dalziel was referring to quality control procedures that were not applied as planned. These involved comparing the students' exam and school scores. The comparisons were intended to test the accuracy of the scores. These checks had not been conducted because data entry problems at the SQA prevented access to the data, and scores were released later than expected, shortening the amount of time available for checks. (Macdonald, 2000c; The Educational Institute of Scotland, 2000).

By the end of 2000, the SQA had not yet recovered from the year's errors. As the Educational Institute of Scotland (EIS) Committee convened to diagnose and correct what went wrong, the teachers' union was warning that the number of markers could continue to decline (Munro, 2000). Given the large anticipated increases in student candidates for the 2001 school year, this prediction was unsettling. In short, although most of the errors had been corrected, the system that produced the errors had not yet been restructured.

For months, teachers and administrators had warned the SQA of the data flow problems, but authorities repeatedly dismissed these signals.

The SQA's acting chief executive, Bill Morton, delivered a report to the Parliamentary Education Committee describing the causes of the Scottish exam disaster. He dismissed the notion that the crisis was due solely to data and information management troubles. Instead, he cited poor project planning and management and inadequate oversight of quality control procedures (Munro, 2000). In other words, Morton's report identified latent error as the cause of the active errors.

For months, teachers and administrators had warned the SQA of the data flow problems, but authorities repeatedly dismissed these signals (The Educational Institute of Scotland, 2000). Consequently, policy makers ignored the concerns of those with first-hand knowledge of how the system was functioning. One observer recommended,"To avoid a repeat of this year's exams disaster, it is vital that the views of teachers, schools, and parents are heard" (Morrison, 2000, p. 3).

Two years after the Scottish incident, England also experienced a systemic error with its A-level examination system. In this case, the Qualifications and Curriculum Authority (QCA) was accused of pressuring the exam bodies to downgrade student exams in order to reduce the student pass rate (Bright & McVeigh, 2002). Within a month of the issuance of student grades, the teachers' union called for an independent examination of the system, and this was conducted by the former schools inspector, Mike Tomlinson (Tomlinson, 2002). Tomlinson's interim report was released on September 27, 2002, approximately six weeks after student grades were issued. The report noted that amendments to the A-level exam program, made when AS-level exams were introduced in 2000,[27] had necessitated changes in standard-setting and grading procedures. However, in the rush to institute changes, the criteria for grading and methods for linking grades from one year to the next had never been fully translated into practice and had been poorly communicated to teachers, students, and even the exam boards.[28] The events set the stage for confusion in scoring, which affected both teachers' and exam board grades. Uncertainties in grading were coupled with pressure exerted by the QCA not to allow the 2002 pass rate to exceed that of 2001. The upshot was that more than 90,000 student exams had to be re-graded, and 168 students were found to have been unfairly denied placement in their chosen university (Harris & Clark, 2002; Tomlinson, 2002). As in Scotland, these pressures were compounded by greater numbers of students taking exams than anticipated, leaving proctors and markers in short supply, thus placing more stress on the system (Timeline: Edexcel woes, 2002; Goodbye GCESs?, 2001). Moreover, the QCA was given a number of additional signals that grading problems were imminent. After more than 6,000 exams required remarking in 1999, with most resulting in higher scores, exam boards Edexcel and Oxford and Cambridge and RSA (OCR) were warned to improve their service (see Appendix A, #30 for explanation of the improvement required). In an op-ed published in *The Times*, researcher and economist David Lines warned that the diminished contact between the exam boards and the schools, and the increased centralization and commercialization of the process could erode standards in the industry:

> *[The exam policies are] built on the erroneous assumption that external examinations are accurate, fair and efficient, while assessment by teachers is not. This notion has been brought about by the determination of successive governments to centralize and control all aspects of education…. So that [exam boards] have increasingly severed their links with their founding universities and become more*

> *commercially orientated…. So now we have an examinations industry, like the railways, shorn of old standards and values, but required to serve increasing numbers of demanding customers. It is hardly surprising that accidents happen (Lines, 2000, p. 1).*

The SQA and QCA cases demonstrate how latent errors, stemming from the design of the testing program, can cause active errors that become evident in the test results. Confirmation of this link may only be made after a careful analysis of the system, however. In both England and Scotland, independent examinations of the problems were ordered and conducted within a short time. By contrast, in the US, timely, systemic reviews such as these have not been as forthcoming.

In Arizona, for example, the state assessment program (Arizona's Instrument to Measure Standards – AIMS) experienced a series of mishaps. Multiple errors were detected in the 2000 eleventh-grade test booklets, a tenth-grade algebra question was found to be miskeyed in 1999 (see Appendix A, #29 & 33 for both these errors), and the contents of the math and writing tests were adjusted because the tests were found to be too difficult[29] (Kossan, 2000a; Pearce et al., 2001). As a result, the DOE postponed the date by which high school students must pass the state assessment to graduate (Kossan, 2000b; Flannery, 2000).

Critics have asserted that the Arizona assessment program had been developed too hurriedly and that resources to support it were inadequate (Flannery, 2000; Kossan 2000b). One critic of the program was a state newspaper, *The Arizona Republic*, which sued the Arizona DOE because it didn't release any of the AIMS test questions[30] – this despite the fact that in the second test administration, 92% of the sophomores who had to pass it to graduate failed (Sherwood & Pearce, 2000; Pearce, 2000b). Another critic was Tom Haladyna, a national expert on testing. Although he had initially helped Arizona DOE officials with test development, he quit, alleging that,"the process was too hasty; [the Arizona DOE] went about it very fast, much faster than anyone in my business would want to do it"(quoted in Flannery, 2000, p. 2).

Critics have asserted that the Arizona assessment program had been developed too hurriedly and that resources to support it were inadequate.

Our review of news reports identified latent errors in testing across the United States, not just in Arizona. In addition to the equating errors made by Harcourt Educational Measurement in 2002 (see page 22-23 of this report), other problems associated with latent causes have included:

- *Insufficient piloting of test items that lead to spurious test results (see McGraw Hill, NY Regents Appendix A #35 & 49, and Appendix D #7, and NC BOE Appendix A #41)*
- *Test designs that instigate a variety of errors (in addition to latent errors described in Scotland and England, see those in Ontario, Canada Appendix B #21); and in DOE ranking programs in California (Appendix C #7), and Colorado (Appendix C #12)*
- *Time schedules that don't allow for a thorough checking of the results (see Harcourt Appendix A #27, and Appendix C #13; and Measurement, Inc. Appendix A #44).*

Although these errors appear to result from latent causes, a systematic inquiry is needed to confirm the causes. By identifying and correcting these latent causes, active errors may be prevented in the future.

# CONCLUSION

This paper contains a sizable collection of testing erors made in the last twenty-five years. It thus offers testimony to counter the implausible demands of educational policy makers for a single, error-free, accurate, and valid test used with large groups of children for purposes of sorting, selection, and trend-tracking.

Even highly reputable testing contractors that offer customers high-quality products and services produce tests that are susceptible to error.

No company can offer flawless products. Even highly reputable testing contractors that offer customers high-quality products and services produce tests that are susceptible to error. But while a patient dissatisfied with a diagnosis or treatment may seek a second or third opinion, for a child in a New York City school (and in dozens of other states and hundreds of other cities and towns), there is only one opinion that counts – a single test score. If that is in error, a long time may elapse before the mistake is brought to light – if it ever is.

This paper has shown that human error can be, and often is, present in all phases of the testing process. Error can creep into the development of items. It can be made in the setting of a passing score. It can occur in the establishment of norming groups, and it is sometimes found in the scoring of questions.

The decisions that underlie the formation of cut scores and passing scores are largely subjective. Glass (1977) pointed out that the idea of objectively setting cut scores that accurately differentiate between students who know and students who don't know is largely a fantasy – there is no clear distinction and no mathematical or logical support for such an idea in the realm of education testing. He wrote,"If ever there was a psychological-educational concept ill-prepared for mathematical treatment, it is the idea of criterion-referencing [i.e., setting cut scores]"(p. 10). When incongruities surface in exam scores between grades or subject matter tests, or between test scores and actual performance, one way to eliminate the effect of human error is to reexamine the cut-score -setting process. This is rarely done, however, because resetting cut scores would severely interfere with the reporting of trend data.

Measuring trends in achievement is an area of assessment that is laden with complications. The documented struggles experienced by the National Center for Education Statistics (NCES) and Harcourt Educational Measurement testify to the complexity inherent in measuring changes in achievement. Perhaps such measurement requires an assessment program that does only that. The National Center of Educational Statistics carefully tries to avoid even small changes in the NAEP tests, and examines the impact of each change on the test's accuracy. Many state DOEs, however, unlike NCES, are measuring both individual student achievement and aggregate changes in achievement scores with the same test – a test that oftentimes contains very different questions from administration to administration. This practice counters the hard-learned lesson offered by Beaton,"If you want to measure change, do not change the measure"(Beaton et al., 1990, p. 165).

Furthermore, while it is a generally held opinion that consumers should adhere to the advice of the product developers (as is done when installing an infant car seat or when taking medication), the advice of test developers and contractors often goes unheeded in the realm of high-stakes decision-making. The presidents of two major test developers – Harcourt Brace and CTB McGraw Hill – were on record that their tests should not be used as the sole criterion for making high-stakes educational decisions (Myers, 2001; Mathews, 2000a). Yet more than half of the state DOEs are using test results as the basis for important decisions that, perhaps, these tests were not designed to support. In an interview with *The Cape Cod Times*, Eugene Paslov, then Harcourt Brace's president, said that standardized tests like Massachusetts' MCAS exam should not be used as the sole determinant of who graduates from high school and who does not. He stated,"When these tests are used exclusively for graduation, I think that's wrong"(Myers, 2001, p. 1). Massachusetts Board of Education Chairman James Peyser responded,"Obviously [the test contractors'] job is to provide a service, not to make policy. No, we don't need them on board"(quoted in Hayward, 2001a, p. 1).

The systemic problems documented in the United Kingdom show that error can also be introduced through poor management decisions. Such latent errors place stress on the entire system, increasing the probability that mistakes will be made. In both Scotland and England, independent audits were conducted to examine the root of the errors. Attributing errors to faulty management decisions is a difficult process, one usually undertaken only after the entire system has suffered a large shock. Both examples also suggest that the first people to notice latent error may be those educators or service providers who are in a position to see, first-hand, the effects of testing programs. Policy makers should therefore heed their concerns. Unfortunately when mistakes are discovered, the tendency to"name and blame"is strong. Once culpability is determined, the search for other causes of the problem usually ends. Problems that are systemic in nature may be able to masquerade as individual failings indefinitely; or at least until minor fix-ups no longer keep the system running. Common testing conditions in the US (and abroad) that introduce latent error are: instituting testing programs without adequate time to evaluate the operation, rushing the piloting process for test questions, and high-volume testing that is conducted within a short period of time. Of the latter a manager of psychometrics services at NCS once said,"When you speed up processes, you increase the risk of there being errors. You just don't have the time to do the quality control"(Fletcher, 2001, p. 2).

Finally, all of these concerns should be viewed in the context of the testing industry today. Lines (2000) observed that errors are more likely in testing programs with greater degrees of centralization and commercialization, where increased profits can only be realized by increasing market share,"The few producers cannot compete on price, because any price fall will be instantly matched by others …. What competition there is comes through marketing"(p. 1). In Minnesota, Judge Oleisky (Kurvers et al. v. NCS, Inc., 2002) observed that Basic Skills Test errors were caused by NCS' drive to cut costs and raise profits by delivering substandard

service – demonstrating that profits may be increased through methods other than marketing. With the recent passage of President Bush's No Child Left Behind Act (NCLB)[31], the testing industry in the US will become increasingly more centralized and more commercialized. An amplified demand for testing services without an appreciable increase in the number of service providers in the short term will intensify time pressures already experienced by the contractors. At the same time NCLB will heighten the reliance of state DOEs on the few contractors available, creating a situation whereby those who pay for a service become increasingly dependent on one that is more prone to error. Coupling these conditions with the lack of industry oversight creates conditions for a future that is ripe for the proliferation of undetected human error in educational testing.

As this monograph goes to press additional errors have come to our attention that could not be included. These and the errors documented in this report bear strong witness to the unassailable fact that testing, while providing users with useful information, is a fallible technology, one subject to internal and external errors. This fact must always be remembered when using test scores to describe or make decisions about individuals, or groups of students.

# END NOTES

1 For an excellent treatment of the power of numbers in our society see Porter, 1995

2 Such an investigation was conducted in Scotland, where a systemwide inquiry showed that active errors were the direct result of latent (management) error (see Appendix A, #38).

3 A miskeyed item is one in which an incorrect response is coded as correct.

4 A cut score is the point on a test score scale that either separates failing from passing scores, or separates scores into performance levels. For a fuller treatment of the issues, see Horn et al., 2000.

5 NES rebuffed the committee's request for technical information, informing them that they should solicit data from the individual states that purchase their tests (Levinson, 2001). Officials from two of the state-run teacher testing programs "reported their understanding that the requested technical information could not be disclosed to the committee because of restrictions included in their contracts with NES" (Mitchell et al., p. 134). Those overseeing the tests in the states provided very little information and any data offered was so incomplete that it was unusable.

6 Percentile scores are used to compare a student's performance with a nationally-normed sample. Raw test scores must be converted into percentiles using a formula or table. In this error, the raw scores were accurate, but the conversion was not.

7 Kline's words proved prophetic when eight months later McGraw Hill contacted Indiana's Department of Education to acknowledge that a programming error affected Indiana's percentile scores and those in five other states (Klampe, 1999).

8 The other 26,332 students actually scored below the 15th percentile. Of the 8,668 students who actually scored above the 15th percentile, 3,492 either did not attend summer school or failed a second test given in summer school. Dr. Crew allowed all 3,492 to move up to the next grade, unless their parents recommended that they continue to be retained (Archibold, 1999).

9 Although this was not widely publicized in the press, it was the second year that Harcourt Brace misclassified LEP students (Asimov, 1999a). The lack of press coverage of the earlier error suggests that it was either caught quickly or was less significant.

10 On the SAT, 100 points is equal to one standard deviation. One standard deviation from the mean includes the scores of 34% of the population in either direction. Again measuring from the mean, a decrease in one standard deviation would mean that, instead of performing better than 50% of the population, the test-taker would score better than only 16% of the population. This represents a huge difference for the college-bound student.

11 These equating errors are discussed in more detail in Part 3 of this paper.

12 Khorkina was not the only Olympian adversely affected by the error. U.S. national vault champion Elise Ray was ranked 35th on the event after falling twice on the faulty equipment. She reran the event, moving from 35th to 14th place. Ray commented, "I didn't know what was wrong. It looked low to me, but I thought it was my nerves" (Harasta, 2000, p. 1).

13 Test scores may be averaged over three years if this will help in a school's rating. Under a time crunch, the Virginia Department of Education had failed to do so.

14 A norm-referenced test compares individual student performance to a norm-reference group (a nationally representative group of students). Scores for the tested group are then transformed to take on the characteristics of a normal distribution so that approximately 50% of the norm group will score above the mean and 50% below.

15 There are two forms of this test: Forms T and S.

16 See Appendix D for similar cut score problems in New Jersey (#3) and Maryland (#4).

17 Kane and Staiger used "filtered estimates" of school-based achievement scores. These estimates were described as "a combination of the school's own test score, the state average, and the school's test scores from past years, other grades, or other subjects" (2001, p. 14). They were further refined by identifying and subtracting noise variance from the total variance.

18 A block is a designated grouping of test items. The items are grouped together to meet certain test specifications, such as content coverage or time constraints.

19 A reduction in the scores needed to pass four social studies tests also contributed to the higher pass rates in 2002 (Seymour, 2001; Helderman & Keating, 2002).

20 Scaling refers to the process of translating raw scores into a score scale that has certain desired properties. For example, on the MCAS the reported score range is from 200-280 per subject matter test while the number of raw score points is about 40. If students get little or nothing right they do not receive a score of "0"; they are instead issued a "200." Through the use of scaled scores test makers can make the test scores not only easier to interpret, but easier to work with (for example, scaled scores from one test administration to the next may be compared; raw scores cannot).

21 See Haney, 2000, for an examination of how student attrition affected score trends in Texas.

22 Statistically significant gains were reported among some age groups in some subjects since 1971/1973 in reading and math, but none of these large gains were reported within the time period discussed here.

23 Usually a small number of test items are not released because they are used to link the test results from one year to the next.

24 For a discussion of the importance of identifying plausible rival hypotheses, see Bickman, 2000.

25 Teachers mark (grade) the tests.

26 Teachers provided SQA with their predictions for each student's performance on the exams. The SQA then compared these with the students' actual scores.

27 AS-level exams are taken at the beginning of high school or secondary education, A-level exams are taken after.

28 See Appendix A, #51, for a fuller explanation of the error.

29 Interestingly, although some of the difficult math questions were eliminated, student scores on the 2001 AIMS tenth-grade exam did not improve over the 2000 scores. A spokesperson for the Arizona Department of Education admitted, "We never said we were going to make the math test easier" (Pearce et al., 2001, p. 1).

30 When the Arizona DOE finally did release AIMS questions, new contractor McGraw Hill requested that they pay $263,000 for damages, citing a breach of contract wherein the DOE promised not to disclose test items. The contract clause was put in place to allow the state to save money – when questions are released, new ones must be created and field-tested and this results in a more expensive test. McGraw Hill claimed that because the DOE had not paid for the testing company to create new items when they were released, the questions were the property of the testing company (Kossan, 2002).

31 NCLB mandates yearly testing in grades 3-8 in reading and math; since most US schools don't test students that often, the amount of testing will increase dramatically. Public schools will be judged according to their performance on these tests, and scores are expected to increase over time, so the probability of latent error from tracking trends will increase. In addition, because states use different tests, test results will be linked to NAEP in an attempt to establish consistency. The linking process (analogous to equating or bridging) will introduce the possibility of further latent error (*The No Child Left Behind Act*, 2002).

# APPENDIX A:
# Testing Errors NOT Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (1) 1981 | PSAT/ETS | 17 year-old test taker from Florida | A Florida student challenged the keyed answer to a question about a pyramid. ETS found the keyed answer was wrong and determined that this student and a quarter of a million others gave the correct answer (Fiske, 1981a & d; Pyramids, Serpents; The exposed side, 1981). | When the error became public, hundreds of people wrote to ETS with their versions of the correct answer. Apparently, the only incorrect answer was the one the test required. ETS did not reduce the scores of those who chose the keyed answer because the results were tied to a National Merit Scholarship and they didn't want to penalize students unnecessarily (Fiske, 1981b). Disclosure of the error was made possible by a 1981 ETS policy that extended benefits from New York's "truth-in-testing" law to examinees in other states (see below) (Fiske, 1981a). |
| (2) 1981 | SAT/ETS | High school senior from New York City | A student who had seen his SAT questions and answers (by virtue of New York's "truth-in testing" law) asked ETS why they included a question with two correct answers. Students had to identify from four rows of numbers the row that contained the square and cube of two different integers. One row showed a 9 and an 8 – the square of 3 and cube of 2 – and the next showed an 8 and a 4, the cube of 2 and square of –2. ETS had not anticipated students coming up with a negative integer (Begley & Carey, 1981; Fiske, 1981c; Walsh, 1981). | ETS corrected the scores of 19,000 students who gave the (unanticipated) correct reply. Their scores increased by 30 points or less (None of the above, 1981; Fiske, 1981c). Though ETS officials acknowledged that the discovery of the error was linked to New York state's "truth-in-testing" law, New York continues to be the only state in the country with such a law (NY CLS Educ. B 342, Article 7-A, 2001; Fiske, 1981c & d). |
| (3) 1985 | Alabama Initial Teacher Certification Test/NES | Team of psychometricians and testing experts from Boston College working for plaintiffs in a lawsuit | Miskeyed items and items of poor technical quality were discovered in eight administrations of the Alabama Initial Teacher Certification Test (from 1981 to 1985). Through the discovery process for the trial of Allen et al. v. Alabama State Board of Education (1999), the plaintiffs found at least 355 candidates who failed the exam because of at least one miskeyed item (Ludlow, 2001). | In a subsequent trial, Richardson v. Lamar County Bd. of Educ. (1989), plaintiff Richardson was awarded damages that included back pay and compensation for lost employee benefits. Other candidates who incorrectly failed the test were given lifetime teaching certificates. The Court's ruling made it unlikely that Alabama would regain full independence in teacher certification until 2015 (Ludlow, 2001; Staff, 1999). |
| (4) 1985-1987 | New York Bar Exam/ State of New York | Applicant/ test-taker | Ambiguous questions on the New York State Bar Exam resulted in at least 137 test-takers erroneously failing the exam. | One hundred eleven failing grades on that exam were withdrawn in 1985 when a test-taker challenged his score; 26 failing grades were revoked on the same exam in 1987 (New York Bar Exam, 1988). |
| (5) 1987-1988 | Tennessee High School Proficiency Test/ Tennessee DOE | Tennessee high school teacher | A high school teacher challenged two ambiguous questions on the Tennessee High School Proficiency Test. On both, students who gave answers that were correct but not the keyed responses failed the exam by one point. | The Tennessee Department of Proficiency Testing opposed giving credit to one correct, but not keyed, response on the grounds that most students chose the keyed answer (J. Anderson, personal communication, November 28, 1987; Dickie, 1987; G. Madaus, personal communication, December 8, 1987). No student scores were changed as a result of the challenge. |

# APPENDIX A:

## Testing Errors NOT Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (6) 1990 | SAT II Chemistry Test/ ETS | Test takers | An SAT II chemistry exam was missing a page of charts necessary for answering 13 of 85 questions; 130 students nationwide took the flawed exam. | ETS notified students that they had to retake the exams if they wanted a score. Those who did not received a $21.00 refund (O'Brien, 1990). |
| (7) 1991 | California Test of Basic Skills (CTBS)/ CTB/ McGraw Hill | A Brookline, MA superintendent and Walt Haney, testing expert from Boston College | A Brookline, Massachusetts superintendent noticed that students with similar raw scores were receiving very different local percentile rankings on the CTBS. Walt Haney corroborated these discrepancies and found that the errors had existed for at least six years. | Haney cautioned the superintendent not to use the local percentile rankings. The testing company took no action, claiming that the discrepancies occurred because the tests were submitted to the company at different times, not all at one time as recommended (W. Haney, personal communication, May 15, 1991). |
| (8) 1993 | High school proficiency exams/ New Jersey DOE | | A New Jersey DOE official was reported saying, "This is an election year – I don't want any bad news," when told that 4,800 students were incorrectly classified as needing remedial teaching according to the results of the eighth-grade and high school proficiency exams (State blames official, 1993). | A report on the incident blamed the problem on poor test supervision and a lack of financing. |
| (9) 1994 | California Learning Assessment System (CLAS)/ CTB/McGraw Hill and the California DOE | Officials from the Orange County School District | Several California school districts scored lower than warranted on a CLAS test when missing tests were scored as zero (it is unclear what happened to the tests). The ratings were questioned when two schools that historically scored high – Corona del Mar and Newhart – were reported to be among the lowest-ranked schools in Orange County on the math test (Wilogren, 1994). | McGraw Hill blamed the error partly on DOE officials who, when asked what should be done, suggested that either zeros or the mean score could be substituted for the missing tests. DOE official Gerry Shelton countered that the DOE never suggested the use of zeros but did suggest substituting the mean score as the best available estimate of student scores. In the end, a computer glitch at McGraw Hill was blamed for the use of zeros. The error resulted in inaccurate district scores as well (Wilogren, 1994). |
| (10) 1994 | California Learning Assessment System (CLAS)/ CTB/McGraw Hill and the California DOE | *Los Angeles Times* | A computer analysis performed by the *Los Angeles Times* revealed numerous scoring and management errors on CLAS assessments from 1993 to 1994. This program tested a mandated minimum of 25% of students from each school, and reported scores in the form of school rankings. The analysis found at least 146 schools with scores for fewer than 25% of students, which resulted in inaccurate ratings for some schools and districts. In some cases, only the lowest-scoring students were tested (Heads should roll, 1994; Wilogren & O'Reilly, 1994). | The *Times* found that only four of 169 tests were used for one elementary school's rating, and that these four represented the lowest mathematics test scores in that school. Although the error affected school ratings only, a *Times* editorial focused on the damage done to schools; "These were not harmless errors. The results prompted parents to ask private schools about getting their children admitted; there were renewed calls to break up the Los Angeles Unified School District" (Heads should roll, 1994, p. 1). The problems were attributed to McGraw Hill's handling of test materials: "answer sheets getting lost…, arriving… defaced, or being split into groups that caused processors to believe that fewer students at certain schools took the tests" (Wilogren & O'Reilly, 1994, p. 2). A panel of statistical experts that included Lee Cronbach of Stanford University concluded that the 1993 school-level scores were unreliable (Merl, 1994). |

# APPENDIX A:

## Testing Errors NOT Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (11) 1995 | Kentucky Instructional Results Information System (KIRIS)/ Advanced Systems in Measurement and Evaluation, Inc. | Employee at Kentucky's DOE | In 1997, officials from Kentucky's DOE asked Advanced Systems for the formula used in scoring the 1995 eighth-grade vocational studies test. The contractor could not provide it. | The contractor had to rewrite the formula as well as rerun the scores. The resulting scores were slightly different. DOE officials accepted the new scores as they were, in part, because the original scores had been released and used two years earlier (Harp, 1997a, p. 2). |
| (12) 1996 | Stanford-9 Achievement Tests (SAT-9)/ Harcourt Brace | Philadelphia school district employees | District Superintendent David Hornbeck announced in 1998 that Harcourt Brace had admitted to a scoring error on the SAT-9 dating back to 1996 and detected by district employees in 1997. This error caused two schools to be classified as needing remedial help when their test scores had actually improved. | The schools were removed from the list of schools subject to "drastic action," that included staff transfers. Harcourt was fined $192,000 for this error and a subsequent one made in 1997 (Jones & Mezacappa, 1998; Snyder, 1998). |
| (13) 1997 | SAT/ETS | High school student from New Hampshire | A high school senior, described as a gifted math student, found two solutions to a multiple choice math problem. The answer depended on whether students chose to work with positive or negative integers. | The student e-mailed the problem to ETS within days of taking the exam. ETS, however, took months to respond due to a mix-up with e-mail delivery at the company (Ford, 1997). When ETS corrected the error, the scores of 45,000 other students rose by as much as 30 points. ETS sent corrected test scores to the candidates' prospective colleges and universities so as not to damage their chances for admission (Curran & Drew, 1997; Siemaszko, 1997; Tabor, 1997; Woo, 1997). |
| (14) 1997 | Kentucky Instructional Results Information System (KIRIS)/ Advanced Systems in Measurement and Evaluation, Inc. | Kentucky DOE officials | State education officials contacted Advanced Systems in June, 1997, questioning elementary test scores that appeared too low. The company found a programming error that did, in fact, yield low vocational studies and arts and humanities test scores. Before it was found, the mistake had cost many elementary schools their share of a twenty-million-dollar reward fund. | Kentucky had to pay out an additional two million dollars to schools that had been denied their reward because the state's reward fund was depleted (Harp, 1997b; State's Schools Eager, 1997; Vance, 1997). In 1997, Advanced Systems lost their eight-million-dollar-a-year testing contract with the state (Harp, 1997b). |
| (15) 1997 | Indiana Statewide Testing for Educational Progress (ISTEP+)/ Indiana DOE | Fort Wayne Community Schools super-intendent | In a scoring discrepancy, students with high percentile rankings were classified as requiring remediation, while those with much lower percentile rankings were said to have met state standards. | State education officials attributed the discrepancy to a small number of questions deemed "essential." If students missed a number of these, they were identified as requiring extra help, regardless of how they performed on the rest of the exam (Ross, 1997). |

# APPENDIX A:

## Testing Errors NOT Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (16) 1998 | Stanford-9 Achievement Test (SAT-9)/ Harcourt Brace | A principal at a Philadelphia elementary school | A Philadelphia elementary school principal wondered how a student scoring above the 90th percentile on the SAT-9 could be classified as performing at the "below basic" level. The company attributed the mistake, which resulted in no students attaining proficient or advanced ratings, to "human error." | The error affected all school and student reports at all grade levels in reading, math, and science. The erroneous results were not released. Parents and students received corrected reports within days. Harcourt Brace promised to prevent future errors by adopting more stringent procedures (Snyder, 1998). |
| (17) 1998 | New Standards Reference Exam/ Harcourt Brace | Vermont and Rhode Island education officials | Education officials in two states noticed errors in scoring that affected fourth-, eighth-, and tenth-grade writing composition test scores. The errors were blamed on scorers who veered from anchor papers used to standardize scoring and tended to assign higher scores than they should have. | The revised results showed lower student achievement in writing overall (Allen, 1999; E. Granger, personal communication, February 23, 2000). |
| (18) 1998 | New Standards Reference Exam/ Harcourt Brace | Employees at the Vermont Newspaper, *The Burlington Free Press* | A "cut and paste" error resulted in low numbers of fourth-grade students at one Burlington school being classified as meeting math state standards. | When the error was corrected, the percentage of students at South Burlington Central School who met the state standards went from zero to 46% (Good, 1998). In January 2000, Harcourt agreed to pay $628,000 for a number of errors that occurred from 1998 to 1999 (see item above as well as Table B, #7) (Ring, 2000). |
| (19) 1999 | Washington's Assessment of Student Learning (WASL)/ Riverside Publishing Company | Washington State education officials | Scores on over 400,000 Washington student essays were inflated when scorers gave too many perfect scores for grammar and spelling. These mistakes occurred despite quality control procedures that were designed to prevent them (Houtz, 1999a). | The revised test results showed lower student achievement in writing overall. Riverside Publishing Company agreed to pay the cost to rescore the exams, which was estimated to be $600,000 (Houtz, 1999b). |
| (20) 1999 | TerraNova/ CTB/ McGraw Hill | Indiana District education officials | Indiana education officials questioned McGraw Hill after they noticed a sharp drop in percentile scores on the TerraNova. McGraw Hill then found an error that stemmed from their use of the wrong norming table (Brunts, 1999; King, 1999). | "Corrected" scores were sent out to parents and students; the results, however, underwent a further correction when another error was found later in the year (ISTEP+ graduation test, 1999; Klampe, 1999). |
| (21) 1999 | TerraNova/ CTB/McGraw Hill | Statistician working for the Tennessee DOE | Statistician William Sandler questioned McGraw Hill after noticing a dip in 2/3 of Tennessee's sample tests' percentile scores. McGraw Hill attributed the error to their use of the wrong norming table. | Only corrected scores were sent out to parents and students. Sandler advised McGraw Hill to check the percentile rankings for the other states as well (Zoll, 1999a & 1999b). |

# APPENDIX A:

## Testing Errors NOT Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (22) 1999 | TerraNova/ CTB/ McGraw Hill | Education officials from Indiana and Tennessee | New York City education officials questioned McGraw Hill about the accuracy of the percentile scores on their city achievement tests because they were unexpectedly low (Hartocollis, 1999b). When officials were assured that scores were accurate, they required students scoring below the 15th percentile to attend summer school or be retained in grade (Hartocollis, 1999c; Archibald, 1999). Because of an error affecting percentile rankings at the low and high end of the score continuum, a quarter of the 35,000 students attending summer school should not have been required to do so (Hartocollis, 1999c). In addition, Dr. Rudy Crew, then Chancellor of Schools, "removed five superintendents and put four more on probation, again citing low scores" (Hartocollis, 1999d, p. 2). In eight out of nine of those schools, scores had actually risen. When corrected, school scores were four percentage points higher on average (Archibold, 1999; Hartocollis, 1999d). | The *New York Daily News* estimated the expense of erroneously sending children to summer school to be $3.8 million dollars. Deputy Schools Chancellor Harry Spence agreed (Gendar, 1999). Rudy Crew recommended that NYC schools continue their contract with McGraw Hill, but that the company be fined $500,000 and submit to an independent audit (Crew Backs Company, 1999; Mendoza, 1999). |
| (23) 1999 | TerraNova/ CTB/ McGraw Hill | Officials from Indiana and Tennessee | McGraw Hill notified Indiana education officials that a second error involving percentile scores had been detected (ISTEP+ faces, 1999; Viadero, 1999; Klampe, 1999; Edelman & Graham, 1999). | The Indiana DOE set up a team to audit the contractor's procedures (Smith, 1999; Gruss, 1999). |
| (24) 1999 | TerraNova/ CTB/ McGraw Hill | Officials from Indiana and Tennessee | McGraw Hill notified education officials in Wisconsin that the same error involving percentile scores in Tennessee, NYC, and Indiana affected thousands of scores on the Wisconsin state achievement test (Thompson, 1999). | Corrected percentile scores showed better student performance than previously reported (Murphy, 1999; Thompson, 1999). |
| (25) 1999 | TerraNova/ CTB/McGraw Hill | Officials from Indiana and Tennessee | McGraw Hill notified Nevada education officials that the state percentile scores were affected by the same error on the TerraNova as in Indiana, Tennessee, NYC, and Wisconsin (Bach, 1999a). | Since Nevada used the test to identify schools with poor performance, three schools were incorrectly classified as inadequate for the eleven months before the error was discovered. All of them were allowed to keep school improvement funds, given to schools so classified (Bach, 1999b). |
| (26) 1999 | Indiana Statewide Testing for Educational Progress (ISTEP+)/ CTB/ McGraw Hill | Fort Wayne Community Schools Administrators | Administrators from Fort Wayne and other school districts noticed a large number of "undetermined scores" in Indiana's high school exit exam. These are scores of students who do not complete the test; most of them also failed the exam. McGraw Hill rescored the tests and found that most of them had been completed. A computer programming error was blamed. | Thirteen students who had been told they had failed the exam had actually passed it (Klampe, 2000). |

# APPENDIX A:

## Testing Errors NOT Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (27) 1999 | SAT-9/ Harcourt Brace | California district officials | Harcourt Brace erroneously classified newly English-proficient students as being "limited English proficient" (LEP); which inflated the aggregate scores for LEP students. Early press reports cited the large gains as evidence that California's Proposition 227 (which decreased the numbers of LEP students in bilingual education) worked (Sahagun, 1999; Colvin & Smith, 1999). | Revised scores indicated much smaller gains for LEP students than originally reported. The average gains of 2-3 percentage points were consistent with expected gains from the use of the same test for two years in a row (Sahagun, 1999; Moran, 2000). The error muddied the evaluation of the effects of Proposition 227 (Mora, 1999; Sahagun, 1999). The California DOE fined Harcourt Brace $1.1 million dollars for this and another error (see #28 below) (Gledhill, 1999). |
| (28) 1999 | SAT-9/ Harcourt Brace | Long Beach, California, employees | The national percentile rankings for 44 of California's year-round schools were miscalculated because Harcourt erred in counting the number of days of school attendance. Because year-round schools had been in session for fewer days than schools with traditional schedules, their reported percentile scores were lower than they should have been (Colvin, 1999b; Moran, 1999). | The error in year-round school results caused a few weeks' delay in issuing test scores. |
| (29) 1999 | AIMS/ National Computer Systems | Arizona State educators | Arizona state educators found a miskeyed mathematics item in the tenth-grade AIMS math test. | After correction, 27% of the scores increased, and 142 more students passed than was originally reported (Pearce, 2000a; Arizona Department of Education, 2000). |
| (30) 1999 | A-level and GCSE Exams/ Edexcel, Oxford and Cambridge RSA (OCR), and The Assessment and Qualifications Alliance (AQA) | Students | Hand-scoring on 1999 A-level and GCSE exams in England, Wales, and Northern Ireland was found to be faulty for thousands of students. Six thousand grades were increased after students challenged their initial scores (Clare, 2000). | The Qualifications and Curriculum Authority ordered Edexcel and OCR to improve after they failed to re-grade most of the exams within 30-40 days, thereby jeopardizing students' university placements (Cassidy, 2000). |
| (31) 2000 | Missouri Assessment Program (MAP)/ CTB McGraw Hill and the Missouri DOE | Ladue School District officials | After the district asked McGraw Hill to rescore 200 essays that they believed received too low a grade, 33 of them received higher grades (Franck & Hacker, 2000). By agreement, school districts could ask to have tests rescored whenever they believed that a sufficient number of them were scored incorrectly. Unfortunately, the $30.00 fee for rescoring (if the new score was the same or lower) was beyond the reach of many poorer school districts (Margin of Error, 2000). | This error focused attention on the Missouri DOE's policy of having essays scored by one person only. In most cases, two or more readers score each essay to standardize scoring and improve reliability. Results from the MAP exams were used to rank Missouri school districts and for school accreditation. Poor school-wide test results, whether accurate or not, may have resulted in schools losing their accreditation (Franck & Hacker, 2000). |

# APPENDIX A:

## Testing Errors NOT Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (32) 2000 | Oregon Statewide Assessment/ Oregon DOE | Oregon educators and students | At least five errors were found on the Oregon Statewide Assessment tests. Mistakes included repeated questions, improperly labeled diagrams, and other printing errors. | The Oregon DOE developed its own tests and therefore corrected its own mistakes (Ore. admits to math mistakes, 2000). |
| (33) 2000 | AIMS/ National Computing Systems | Students, teachers | Arizona students and teachers found various errors on the eleventh-grade AIMS exam. Among these were grammatical errors, misleading or poorly worded questions, and math questions with either no correct answer or several correct answers. | The alleged errors went uncorrected; state DOE officials said corrections would be made only if the aggregate testing data appeared "flawed" (Pearce, 2000c, p. 1). |
| (34) 2000 | MCAS/ Harcourt Brace | Local educators | Several Massachusetts educators detected a printing error on one-sixth of the eighth-grade MCAS science tests. The error prompted students to finish before the test concluded. | Officials from the Massachusetts DOE said that students would not be penalized for incomplete exams caused by this error (Coleman, 2000). |
| (35) 2000 | TerraNova/ CTB/ McGraw Hill | Director of Testing, NYC | In 2001, NYC's then director of testing, Robert Tobias, accused contractor McGraw Hill of delivering inflated scores to 60,000 students in 2000; he suspected that the company had overestimated the difficulty of the new items in the city's sixth-grade reading test for that year, giving students more credit than they deserved (Campanile, 2001; Goodnough, 2001; Kowal, 2001). | The contractor reviewed the scores from 2000, but could find nothing wrong. Company president David Taggart was quoted as saying, "It looks anomalous. Does that mean those scores were wrong? No. It means those students had a good year last year in terms of their performance on the test" (Goodnough, 2001, p. 1). Sixty thousand of 73,800 scores were affected. In one newspaper account, Robert Tobias reportedly estimated that thousands of students who should have been retained in grade were incorrectly promoted because of the inflated scores (Campanile, 2001). In another report, however, Tobias was said to have been sure that no students were improperly promoted because other factors entered into promotion decisions (Gewertz, 2001). |
| (36) 2000 | Basic Standards Test (BST) (Form B)/ National Computing Systems | A Minnesota parent | A parent's persistence paid off when he was able to examine the BST that his daughter had failed. It took this attorney parent two months of sustained effort to view the test. He and the DOE found errors that caused 7,930 students to be incorrectly informed that they had failed. This mistake occurred when the answer key for Form A was used for Form B. Fifty of these students were seniors who were denied a high school diploma, and some of them were not allowed to participate in their high school graduation ceremonies (Children, Families & Learning, 2000a; Drew, Smetanka & Shah, 2000; Carlson, 2000; Drew & Draper, 2001). The parent had flexible working hours that allowed him to pursue the problem. In an interview with the *Star Tribune* he asked, "What if you're a 9-to-5 employee, or… a single parent, or an immigrant? There's no way you'd have ever made it through" (Grow, 2000, p. 1). Another, smaller error involving a question with a design flaw was reported on the Form B math test. With this, 59 students who were told they had failed actually passed (Children, Families, & Learning, 2000a). | In a press conference on July 23, 2000, the State Education Commissioner promised to release corrected scores by summer's end. She further required the test contractor to apologize publicly to the citizens of Minnesota as well as submit to an audit at the contractor's cost. NCS offered to provide a $1,000 scholarship to each senior who was wrongly denied a diploma (Bowman, 2000). In another attempt to make amends, Governor Jesse Ventura handed out certificates to those he termed "innocent victims" (Bakst, 2000, p. 1). Minnesota's Department of Children, Families and Learning established a quality control office for Minnesota's state test following this error. |

# APPENDIX A:

## Testing Errors NOT Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (37) 2000 | MCAS/ Harcourt Brace | Local school administrators | Massachusetts school officials notified Harcourt Brace when some test scores were missing in their school results. Harcourt searched for the missing tests and found most of them. | Tests in one town, Chatham, had still not been found a year later (Myers, 2000; Vaishnav, 2000). Then president of Harcourt Brace, Eugene Paslov, visited Chatham in the spring of 2000 to apologize (Myers, 2001). |
| (38) 2000 | Scottish Qualification (Higher) exams/ Scottish Qualifications Authority (SQA) | Students and teachers | Thousands of Scottish secondary students were given wrong or incomplete marks on secondary school exit exams, resulting in hundreds of students who were denied their first choice in university enrollments (Macdonald, 2000b). The problems occurred after changes were made to the exam program that included forming the SQA three years prior (Macdonald, 2000a) and a change in 2000 that allowed students to take exams later in the year. | The scope of the errors caused the head of the SQA to resign (Clarke, 2000), and instigated a system-wide investigation of the problem. A report listed several causes that included: (a) the schedule for grading exams was too tight, (b) there were more exams taken than anticipated, (c) a system for comparing students' exam grades to school grades was not functioning properly (Macdonald, 2000c), (d) teachers' concerns were overlooked (Teachers' exams fears, 2000; The Educational Institute of Scotland, 2000), and (e) poor project management (Post-mortem, 2000; MacBride, 2000). Two weeks after the report was released, the education minister resigned (Wormersley, 2000). |
| (39) 2001 | MCAS/ Harcourt Brace | A Mass-achusetts tenth-grade student and fourth grade students | A sophomore taking the MCAS found an error on a math question, which asked students to pick out a shape that could not be created by joining equilateral triangles . The student realized that all of the shapes listed could be made by joining the triangles and then chose "polygon" as the answer. In fact, a polygon could be made in that way, only a "regular" polygon could not, but the word "regular" had been omitted (Lindsay, 2001). A smaller error on the fourth-grade English/Language Arts exam identified President James Madison as John Madison (Myers, 2001; Vaishnav, 2001). | Massachusetts Education Commissioner, David Driscoll said that if questions were flawed, they would be removed from the test. He called the test flaw a minor one, saying that, "It's [the test] not going to be perfect. It needs to be close to perfect," (Lindsay, 2001, p. 2). |
| (40) 2001 | Basic Standards Test, Form A/ National Computer Systems (NCS) | A Department of Children, Families and Learning employee | An employee of Minnesota's Department of Children, Families and Learning found a small typographical error in the answer to a question (Draper, 2001). | The question did not count toward students' scores because it was a pilot question. |

# APPENDIX A:

## Testing Errors NOT Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (41) 2001 | ABCs of Public Education/ North Carolina State Board of Education | Employees of local schools | On North Carolina's 3-8 state tests, officials changed the math tests and the passing score, but were unable to perform the amount of field testing necessary to ensure that the new test results were compatible with the old ones. As a result, students passed the new math tests at remarkably high rates that at some schools exceeded 95% (Silberman, 2001). State officials decided not to change the pass rate and to let students keep their scores. | This error coincided with a state policy mandating that children who fail the year-end tests be retained in grade. Since most students passed the math test, most were promoted. BOE officials promised to adjust the following year's passing rate, and so it was anticipated that more students would be retained in grade then (Silberman, 2001). A similar error was reported in a *Washington Post* article: all fifth-grade students were promoted because the pass rate on that test was set too low (Fletcher, 2001). |
| (42) 2001 | Maryland Writing Test/ Maryland State DOE and Measurement, Inc. | Staff at *The Washington Post* | An investigation by *The Washington Post* questioned the scoring on Maryland's writing test for middle-school students. The *Post* found that inconsistent scoring criteria for student essays caused some poor essays to receive passing scores while other, better essays were scored as "failing." In particular, some essays that passed were filled with grammatical and spelling errors as well as poor content, while some that failed showed far better mastery of writing mechanics and content. A state testing official was unable to explain or justify some of the scoring (Perlstein, 2001). | On the basis of the state test scores, the Howard County Board of Education decided to either retain failing students or to provide them with remedial help in summer school. In 2001, 95% of the students faced retention because of test scores (ibid.). |
| (43) 2001 | A-level physics exam/ Assessment and Qualifications Alliance (AQA) | UK students | British students taking an A-levels physics exam found a question that could not be answered with the information provided. The problem required students to calculate the moon's gravitational force, but failed to give the radius of the moon. Three thousand students were affected by the error (Woodward, 2001). | It was reported that students spent dozens of minutes on the question before realizing that it could not be answered. The AQA apologized for the error, but was unsure of its effect on students' final scores. |
| (44) 2001 | New York Regents Mathematics Exam/ Measurement Inc. | Local educators and students | New York high school students and teachers complained about typographical errors in a math retest students had to take in order to graduate. One question was so flawed that it had to be thrown out. Alfred Posamentier, a mathematics professor at City College, expressed concern that bright students would have been the most negatively affected by the errors (Hartocollis, 2001). | The testing contractor blamed the errors on the tight exam schedule (ibid.). |
| (45) 2002 | MCAS/ Harcourt Brace | Science professor at UMASS/ Boston | A science professor detailed errors from four items on the 2001 tenth-grade MCAS exam. Two of the items were so flawed that he believed they should have been dropped from scoring; for another, there was more than one correct answer; and in yet another, a boxplot was used incorrectly, inviting confusion (Gallagher, 2001). | All items were retained in the 2001 scores. |

# APPENDIX A:

## Testing Errors NOT Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (46) 2002 | AS-level government and politics exam/ Edexcel | UK students | An exam error that misstated the number of MPs elected between 1997 and 2001 caused some students not to select that essay question and, instead, to respond to another. Some educators observed that the flawed question prevented students from demonstrating their competency on a topic they had spent months preparing for (Peachey, 2002; Exam board admits 'printing error,' 2002). | Edexcel blamed the error on the printers (Timeline: Edexcel woes, 2002). |
| (47) 2002 | Standards of Learning (SOLs)/ Harcourt Brace | Employees from different school districts | Asked to review the writing test scores by the Virginia DOE, Harcourt determined that its new equating program set the cut score one point too high (Akin, 2002). When the cut score was lowered, 5,625 fifth- eighth- and high school students who had "failed", passed; and an additional 7,702 had their scores raised (Samuels, 2002). This was Harcourt's third equating error of the year (See Appendix B, #22 & 24 for the other two errors). | Since the SOLs affect state-sponsored school accreditation, some speculated that more schools would earn accreditation when the scores were corrected (King & White, 2002). |
| (48) 2002 | STAR/ California DOE & ETS | Local educators | Students in dozens of California schools were confused by the essay directions on the STAR fourth- and seventh-grade writing tests. Most were fourth-graders who responded to the cover directions to open their booklets and write a story, but failed to notice further directions inside the cover that told them what to write about (Magee, 2002). | School officials were concerned that the confusion over directions would lower their state rankings. State testing personnel dismissed those fears, believing most students had followed the directions (ibid.). |
| (49) 2002 | TerraNova/ CTB McGraw Hill | NYC BOE | Plummeting scores on the city's seventh-grade TerraNova reading tests caused the NYC BOE to delay release of the results until a review of the scoring process had been conducted (Gendar, 2002). This error seems to be linked to the 2001 sixth-grade error in which reading test scores appeared too high (see #35). | McGraw Hill officials maintained that the scores were correct, but consented to the review. In the end, they insisted that the scores were correct; but they were not released to the public. In September 2002, the NYC school system terminated its contract with McGraw Hill and hired Harcourt Educational Measurement instead. The decision to go with Harcourt was not a difficult one as the NYC BOE, "sent a request for proposals to 15 testing companies [and] only two, CTB McGraw Hill and Harcourt, had responded" (Goodnough, 2002, p. 1). |

# APPENDIX A:

## Testing Errors NOT Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (50) 2002 | MCAS/ Hartcourt Educational Measurement | A social studies teacher | Teacher John Gibbons identified two correct answers on a multiple choice item in the MCAS eighth-grade history test, only one of which was keyed as correct. Gibbons contacted the DOE and received no response. The DOE did credit students who chose the unkeyed, correct response. Thus, 666 eighth-graders passed after being told they had failed, and 883 other students moved up one performance category. In all, 14,000 students out of 75,000 had been given scores one point too low (Vaishnav, 2002; Nugent, 2002). | The DOE issued new scores for students who went up one performance category. They did not notify students whose scores increased, but had not changed performance levels (Vaishnav, 2002). |
| (51) 2002 | A-level Exams, England/ Oxford and Cambridge and RSA (OCR), Assessment and Qualifications Authority (AQA), and Edexcel | Local teachers, students, and parents; Investigative report published by *The Observer* (Bright & McVeigh, 2002) | England's exam boards, particularly OCR, were accused of lowering exam grades in response to pressure from the Qualifications and Curriculum Authority (QCA) to maintain the A-level's rigorous standards. The QCA's concern began with the high pass rate of 2001; when scores in 2002 were even higher, QCA advised the marking agencies to bring scores into line with 2001 (Timeline: A-level grading row, 2002; Miles, 2002). Students and teachers first became aware of the markdowns when the grades came in. Some students who had always received straight A's in their courses and on most of the A-level tests received "unclassified" (failing) grades on one or two of these tests, resulting in much lower average grades than expected and losses of desired university placements. The disparity in students' scores caused teachers to accuse the various exam boards of grade fixing (Bright & McVeigh, 2002). The teachers' union demanded exams be re-scored (Hayes, 2002). Although a report issued by the QCA first blamed the problem on "poor teaching" (Harrison, 2002), the agency soon after agreed to re-score more than 90,000 exams (Harris & Clark, 2002). | Of the 91,000 students affected, 1,945 had their grades increased and 168 were found to have been wrongly denied admission to their university of choice (Harris & Clark, 2002). Some students who lost coveted university placements joined to sue the UK government and exam boards (Sheldon, 2002) for up to £50,000 each. The head of the QCA, Sir William Stubbs, was fired although neither he nor any other person or organization admitted responsibility for the incident (Harris & Clark, 2002). Sir William threatened to "sue the government for wrongful dismissal unless… a public apology" was given (Timeline: A-level grading row, 2002, p. 4). A report issued by Tomlinson (2002) attributed the problem to a systemic failure; a change in grading the A-level exams was never fully implemented or explained and caused confusion.[1] |
| (52) 2002 | Massachusetts Comprehensive Assessment System retest (MCAS)/ Harcourt Educational Measurement | A high school student | A high school student who used a spatial instead of a numeric solution for a math multiple choice problem based on the binary system identified a second correct response on an MCAS retest. Upon "conferring with mathematicians" (Massachusetts Department of Education, 2002b, p. 1), the DOE agreed that the student's solution was correct and increased the scores of other students who also chose this answer. | Because this retest was taken by students who had already failed the high school graduation test, 449 juniors and seniors who were not eligible for graduation had now "earned a competency determination" (p. 1), thereby allowing them to receive a high school diploma. The senior who found the alternate solution earned a score of 218 on this test, so she was still ineligible to graduate (Kurtz & Vaishnav, 2002). Her cleverness was praised, however, by the state's Education Commissioner, "This girl was able to take a typical math question and come up with a completely unique method of solving it that even our math experts… never considered" (p. 1). |

[1] *Specifically, the report attributed errors to the transition from the previous A-levels to the new AS- and A2-levels. AS-levels, taken by students at the beginning of secondary school, were to be less rigorous than A2-levels, taken toward the end of secondary school. The different levels required that (a) the standards for course work for each be different and (b) the grades from each level type be weighted differently, so that when the grades were aggregated, those from the more demanding A2-level courses influenced the final grade more than those from the AS-level courses. According to Tomlinson, these differences were never made clear, nor were the statistical methods for grade aggregation fully explored or understood. Evidence was presented that showed the three exam marking boards (OCR, AQA, and Edexcel) applied different criteria for marking and aggregating the exam papers (Tomlinson, 2002; Hayes & Linden, 2002).*

# APPENDIX B:
# Testing Errors Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (1) 1976 – 1980 | US Armed Services Vocational Aptitude Battery (USVAB)/ US Department of Defense | US Department of Defense | An undetected calibration error resulted in the enlistment of 300,000 armed services recruits who would otherwise have been declared ineligible due to low test scores. | Overall performance of these recruits was slightly below that of their test-eligible peers. However, many of them performed as well as or better than their peers (Sticht, 1988). |
| (2) 1979 | Maryland Bar Examination/ ETS | ETS | Two months after a Georgetown University Law Center graduate was told he had failed the Maryland Bar Exam, ETS notified him that he had passed the test. A computer scanner that crinkled the exam papers was blamed for the error that affected 59 applicants (Kiernan, 1979). | Only one person's score went from failing to passing. |
| (3) 1980 | SAT/ETS | ETS | Six weeks after sitting for exams, ETS informed 163 Montgomery County seniors that their tests were lost (Brown, 1980). | The students had to re-take the exams. ETS attached a letter to these scores to explain their delay. |
| (4) 1981 | SAT/ETS | ETS | Approximately 1,000 California students were required to retake the SAT after ETS informed them that their test scores would not count because some test items appeared more than once on the test. | Students who chose not to retake the three-hour exam were offered a refund of $20.00 (Eng, 1991; ERROR will force 1,000, 1991). |
| (5) 1992 | Connecticut Mastery Test/ Connecticut DOE | Connecticut DOE | After the DOE determined student scores on the essay portion of the Connecticut Mastery Test were too low, it arranged to have 75,000 sixth- and eighth-grade essays rescored. The DOE then determined that the second set of scores were too high (Frahm, 1992). | Chester Finn Jr., in a statement made to the *Hartford Courant*, suggested that part of the problem lay in the subjective nature of scoring open-ended test items. He said, "The more subjective the [test], the greater the risk of a glitch" (ibid, 1992, p. 2). |
| (6) 1994 | Connecticut Academic Performance Test (CAPT)/ Harcourt Brace | Harcourt Brace | Harcourt Brace was fined $85 thousand, the maximum penalty allowed, by the DOE for sending out the wrong CAPT scores for grades four, six, eight, and ten (State Fines Company for Test Errors, 1994). | The incorrect tests were sent back to the contractor before they were distributed to students. |
| (7) 1998 | New Standards Reference Examination/ Harcourt Brace | Harcourt Brace | Two miskeyed items were found on the tenth-grade writing conventions portion of Vermont's New Standards Reference Examination in English Language Arts (E. Granger, personal communication, February 23, 2000; Sutoski, 1999). | |

# APPENDIX B:

## Testing Errors Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (8) 1997-1998 | SATII/ ETS | ETS | An error in grading caused scores on the mathematics IIC, Japanese reading, and listening tests to be too high; some by as few as 20 points (1/8 of a standard deviation), while others were inflated by 100 points (one standard deviation) (Sandham, 1998). | Four months after sitting for the exam, the 15,500 students affected were told that their scores would drop by an average of 20 points (Weiss, 1998). |
| (9) 1998 | Missouri Assessment Program (MAP)/ CTB/McGraw Hill | McGraw Hill | A calculation error resulted in incorrect scores on the MAP tests at grades four, eight, and ten. | This error had a positive effect on students in low-scoring schools, where overall scores increased, and a negative effect on students of high-scoring schools, where overall scores decreased (Bower, 1998). |
| (10) 1998 | Missouri Assessment Program (MAP)/ CTB/McGraw Hill | CTB/ McGraw Hill | The second 1998 Missouri error occurred when McGraw Hill misreported disaggregated group scores on about 35% of the MAP district score reports (Singer, 1998). | The affected districts were notified of the mistake and McGraw Hill released corrected reports. |
| (11) 1998 | Florida Comprehensive Assessment Test (FCAT)/ CTB/McGraw Hill | McGraw Hill | "An errant computer scanner" that counted all responses marked B as incorrect was blamed for an error that affected about 19,500 of 650,000 test-takers. Errors were discovered on both the verbal and math sections of the tenth-grade test as well as on the math sections of the fifth- and eighth-grade tests. | The corrections yielded higher scores. While these increases were small overall, many students and some schools saw scores go up by as many as 13 or 14 points.[1] The incident was blamed for weakening public confidence in FCAT results (de Vise, 1998a & b). |
| (12) 1999 | SATII/ ETS | ETS | The SATII scores of 1,500 high school students increased by as much as 100 points (one standard deviation) after a mistake was found and corrected. An errant optical scanner that misread ten math questions on the score sheets was blamed for the error (Frahm, 1999). | Robert Schaeffer of FairTest indicated that this error underscores the necessity of "truth in testing" laws that allow students to review the test as well as their answers. |
| (13) 1999 | Delaware Student Testing Program (DSTP)/ Harcourt Brace | Harcourt Brace | A Harcourt employee who used 1998 data to calculate 1999 scores caused an error that affected about four percent of 64,000 test-takers in grades three, five, eight, and ten. The adjusted scores showed that the reading test was passed by more third- and eighth-graders than originally indicated, while fifth-graders passed the math portion in greater numbers. There was also a higher failure rate than earlier reported by eighth- and tenth-graders. (Jackson, 1999). | As of spring, 2000, high stakes were attached to the tests that included grade-level promotion and the issuance of high school diplomas. |

[1] *The 2002 technical manual showed the range of standard deviations from the 2000 FCAT administration for all students to be from 48.03 to 61.97, therefore the 13- to 14-point score difference would represent about 1/4 of a standard deviation on that year's tests (Florida Department of Education, 2002).*

# APPENDIX B:

## Testing Errors Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (14) 2000 | Arizona's Instrument for Measuring Standards (AIMS)/ National Computer Systems | National Computer Systems | Some school-wide scores in writing were skewed when eleventh-grade examinees were identified as sophomores. The adjusted tenth-grade scores increased slightly while eleventh-grade scores decreased slightly. (Pearce & Flannery, 2000). | DOE officials placed blame for the error either on the schools or on the testing contractor. |
| (15) 2000 | Stanford-9 Achievement Tests (SAT-9)/ Harcourt Brace | Harcourt Brace | For the second year in a row, an error emerged in SAT-9 scores for year-round schools. Harcourt Brace immediately acknowledged the problem and the scores were delayed a few days (Note to Readers, 2000). | |
| (16) 2000 | School Certificate Exam/ New Zealand Qualifications Authority | New Zealand Qualifications Authority | Officials found a question on the 2000 School Certificate Math Exam for which there was no correct answer. The question led students to two different answers, depending upon the method they chose to solve it – neither of them was entirely correct, however. | As the weight of the question was 2/3 of a percentage point of the entire exam, officials determined that this mistake would have a negligible effect on the students' total test scores and was not corrected (Larkin, 2000). |
| (17) 2001 | Information Technology Exams/ Edexcel | UK Qualifications and Curriculum Authority | A computer programming error caused more than 10,000 British students to be given the wrong test results on an information technology examination; 3,705 students were told they had failed the examination after having been informed that they had passed it, and another 6,446 who had originally failed the exam were told they had passed it (Students given wrong test results, 2001). | Edexcel apologized for the error, which was discovered in an audit by the Qualifications and Curriculum Authority, vowing, "…to ensure that this never happens again" (ibid, p. 2). |
| (18) 2001 | Graduate Management Admission Test (GMAT)/ ETS | ETS | Approximately 1,000 people who took the GMAT in February and March of 2000 received no credit for nine questions due to a programming error. The average loss was 44 points; however, some test-takers lost as many as 80 (Henriques, 2001). It took ETS six months to publicly announce the error after it was found accidentally by employees conducting "unrelated research" (p. 1). Instead of initially releasing information about the error to the press, ETS decided instead to notify examinees by mail, using the addresses provided at the time of the exam. | Several examinees did not learn of the error on the GMAT until it was publicized, mostly because their addresses had changed since they sat for the exam more than a year before |

# APPENDIX B:

## Testing Errors Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (19) 2001 | Florida High School Competency Test/ Florida DOE | Officials from the Florida DOE | DOE officials had students at Northwestern High School in Miami take a retest when a test booklet was found missing and cheating was suspected. Suspicions of cheating were further bolstered when passing rates at the school rose sharply from 20% on a pre-test given earlier in the year to almost 70% on the actual test. The school principal noted, however, that most scores had increased only a few points – from 698 or 699 on the pretest to just over 700, the passing score, on the actual test (De Valle, 2001). Fifty-nine of those who passed the High School Competency Exam the first time they took it failed it the second, thus making them ineligible to graduate. | Education Commissioner Charlie Crist first decided to allow 17 of the 59 to graduate because their scores were within a few points of the cut-off (Stepp, 2001a). Crist then met with the remaining 42 students, to discuss the DOE's decision to deny them diplomas. He decided to allow the students to graduate with their classmates, in part because the accusations of cheating had not been confirmed, and also because they were given less than two days to prepare for the retest (Stepp, 2001b). |
| (20) 2001 | AIMS/ NCS | Arizona DOE | The DOE held back AIMS writing scores on the 2001 test after observing large incongruities between the 2000 and 2001 scores. An investigation revealed calculation errors in the scores for grades three and five (AIMS writing test scores delayed, 2001; Test scores for some students, 2001). | NCS employed Arizona teachers to, "reset the performance standards used on the writing test in 2000 and … 2001" in an attempt to generate more accurate scores in the future (Test scores for some students, 2001, p. 1). |
| (21) 2001 | Provincewide Tests in Ontario, Canada/ Education Quality and Accountability Office (EQAO) & Moore Corporation Ltd. (scanner) | Officials from EQAO | Scanning that missed whole testing sections from some schools' grade 3, 6, and 9 tests, and missing sixth-grade answer booklets resulted in student and school scores that were too low (Walters, 2001a; Botched test handling, 2001). Due to the loose-knit nature of the testing program in Ontario (many testing services were contracted out to smaller venders with little central oversight) the EQAO was not initially able to identify the schools where tests were lost, nor could they determine the processing error rate in each school. (Marr & Walters, 2001; Walters 2001b) Answer sheets that were scanned at schools in uncontrolled conditions was one of the causes cited for scanner malfunctioning (Walters, 2001a). | A spokesperson for Moore said that the company was overwhelmed by the large number of tests it had to scan (a volume 15% greater than predicted). The company also confirmed that the EQAO exerted little oversight over their scanning work (Marr & Walters, 2001). |
| (22) 2001 & 2002 | Stanford 9 Achievement Tests/ Harcourt Brace | Harcourt Brace | When the 2001 test results were delayed by a month, the president of Harcourt Brace told Georgia's Board of Education (BOE) that there would be no further problems with the test (Salzer & MacDonald, 2001). Instead of having an event-free year in 2002, Harcourt delivered results that deviated substantially from those of the previous year. The results were found to be so riddled with errors and so late that they were deemed to be unusable by the BOE chairwoman, Cathy Henson (Donsky, 2002; State despairs of getting, 2002). Harcourt attributed the errors to problems with equating. | The BOE voted to withhold Harcourt's $600,000 payment. |

# APPENDIX B:

## Testing Errors Detected by Testing Contractors

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (23) 2002 | Higher Exams / Universities and Colleges Admission Service (Ucas) and SQA | Ucas | Scotland's Ucas issued passing grades to students who had failed Higher Exams. A Ucas official explained that the students affected had received scores just under the cut-off, and that the agency had, "given a 'fall back' mark" which was interpreted as a 'pass' (University body 'sorry' for mistake, 2002, p.2). | Ucas issued corrected scores to both the students and to their prospective universities. Education officials raised concerns about an assessment system that created large errors in two out of three years (see Appendix A, #38). SQA launched an investigation to determine the cause of this error. A separate investigation was also undertaken to determine the cause of a 2.2% drop in the number of candidates passing the Higher exams. |
| (24) 2002 | Stanford 9 Achievement Test/ Harcourt Educational Measurement (formerly Harcourt Brace) | Harcourt Brace | Two years after the Nevada state school board terminated its contract with McGraw Hill because of that company's errors; it fined the new contractor, Harcourt Educational Measurement, $425,000 for a testing error that caused 736 students to be told that they failed the high school graduation test when they had actually passed it. The error was attributed to an equating mistake in which the number of questions required to pass the math test was calculated incorrectly (Hendrie & Hurst, 2002). The difference in the number of questions needed to pass the test was one: from 42 to 41 (Ritter, 2002) | The graduation test was taken only by juniors and sophomores, therefore, no students were stopped from graduating. At least one student who failed the exam hired a tutor, however, and Harcourt agreed to pay for the cost of tutors hired for students affected by the error (Vogel, 2002). Several BOE members expressed anger over the incident and one was quoted as saying, "The stuff I want to say you can't print.… I think we should get rid of the company" (Lake, 2002). While the BOE renegotiated with Harcourt for another year, they added a clause to the contract that mandates the immediate removal of the contractor in the event of another large mistake (Hendrie & Hurst, 2002). |
| (25) 2002 | North Carolina's online computer tests for special education students/ North Carolina's Department of Public Instruction (DPI) | DPI | The DPI announced that it was working on a number of problems with the state's new online computer testing program for students enrolled in special education. The new tests allowed the order of questions to be adapted to the ability levels of individuals so that more special education students could be included in the state test program. Due to programming and computer problems, however, hundreds of students had to retake tests and many others had to reschedule exams (Lu, 2002). | |
| (26) 2002 | Colorado State Assessment Program (CSAP)/ The Colorado DOE | DOE staff | A change in directions on an anchor question resulted in the release of inaccurate trend information. Specially, fourth-grade writing proficiency rates showed a decline when scores had actually risen. | Had the error not been caught, many school writing scores would have been too low and schools already graded as "unsatisfactory" could have had sanctions unfairly levied against them (Yettick, 2002). |

# APPENDIX C:
# Errors in School Rankings

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (1) 1989 | California Learning Assessment System (CLAS)/ University of Chicago (sub-contractor for ETS) | California DOE | School rankings for two consecutive years (1987 and 1988) were wrong because a different (and incorrect) computational formula was used each year. The problem was attributed to the development of a formula that excluded the number of students per school (Calvano, 1989). | Each time the errors created erroneous rankings at the low and high ends of the spectrum. Rankings in the middle were found to be correct. Correcting the error would have resulted in revised rankings for an estimated 30% of San Diego Country schools (Calvano, 1989). |
| (2) 2000 | Standards of Learning (SOLs)/ Virginia DOE | School district administrators | Soon after ratings were released in October 2000, administrators from the Virginia Beach school department challenged the rankings of several elementary schools as lower than projected. An omission was discovered in the calculations; it was estimated that dozens of schools were affected (Warchol & Bowers, 2000). | |
| (3) 2000 | TAAS/ Texas Education Agency | Local school administrators | Two schools claimed that data entry errors lowered their school grades. In Fort Worth, the Eagle Mountain-Saginaw schools claimed that the state failed to include the scores of students attending alternative behavioral programs, resulting in a higher dropout rate than was actually merited. (Texas uses dropout rates in conjunction with attendance rates and test scores to grade schools.) Administrators from the Park Cities district in Dallas also alleged that their district earned a lower grade than was warranted. In this case, the state categorized students who transferred to private schools as "dropouts" (Melendez, 2000). | |

# APPENDIX C:
## Errors in School Rankings

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (4) 2000 | Florida Comprehensive Achievement Test (FCAT)/ Florida DOE | Local school administrators | Two elementary school ratings were contested in Florida where schools were graded "A" through "F" on the basis of FCAT scores. The schools had raised their reading scores by the amount necessary to earn higher grades; however, because the DOE rounded the scores down instead of up, school ratings were lower than expected. Schools that improved their ratings by one grade received one hundred dollars or more per student in incentive pay. | One of the schools had already received its bonus money; the other, Lake Myrtle in Pasco, was awarded $95,168 when the error was corrected (Fischer, 2000). |
| (5) 2000 | TerraNova/ New Mexico DOE | State superinten-dent of schools | Days after 94 schools with purportedly large score increases on the TerraNova were promised substantial rewards in the form of school bonuses, the state superintendent of schools announced that the years 1999 and 2000 had been mixed up, resulting in an inaccurate bonus list. Actually, many of the schools on the 2000 list had experienced large drops in standardized scores. The revised list of 101 most-improved schools contained none of the 94 schools named on the first list. | Teachers and principals to whom money was promised and then denied expressed disap-pointment at the reversal. A principal of Albuquerque's Collet Park Elementary School said, "They left us [at] the altar. We had it and now it's gone. We received a little standard-ized fax. It's pretty demoralizing to the staff" (Schoellkopf, 2000, p. 2; Gewertz, 2000). |
| (6) 2000 | Stanford-9 Achievement Test (SAT-9)/ Local School District Computer Programmer | Kenji Hakuta, Stanford University | The *New York Times* ran a front page story about the success of California's Proposition 227 (California legislation that mandated schools to educate limited English proficient, LEP, students in English-only classrooms). The story lauded the remarkable gains in SAT-9 scores in Oceanside School District – a 19 percentage point gain in scores between 1998 and 2000. Oceanside had almost completely dismantled its bilingual programs, servicing most of its 5,000 LEP students in English-only classrooms. These score gains were pitted against a neighboring school district's scores, Vista, with gains that were half those of Oceanside. Vista had granted thousands of waivers to allow students to continue in bilingual classrooms. The explicit message in the *Times'* article was that bilingual programs don't work (Steinberg, 2000). | Among the critics of this analysis was Stanford Professor Kenji Hakuta, who pointed out that score gains between California districts that retained bilingual education were similar to gains made in California districts that had ended these programs (Orr et al., 2000). Then came news that 2,036 of Vista's LEP scores had been omitted in the early analyses, making score comparisons between Vista and any other town suspect (Buchanan, 2000). |

# APPENDIX C:

## Errors in School Rankings

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (7) 2000 | STAR/ California DOE | School Officials in San Diego County and other CA districts | In California, data derived from incorrectly completed questionnaires filled out by young children, resulted in erroneous school rankings. Published school rankings lacked accuracy[1], and bonus money that was to accompany score gains was delayed (Groves, 2000; Spielvogel, 2000). | Corrected school rankings were published months later. State education officials maintained that, in spite of the changes, "…they still [could] not vouch for the accuracy of the new information" (Colvin, 2000, p. 2; Shafer 2000b, p. 1). |
| (8) 2000 | Academic Performance Index (API)/ California DOE & Harcourt Brace | Harcourt Brace | In an ironic reversal of typical practice, Harcourt Brace fined California school districts between $500 and $2,000 for data entry errors made at the district level. These errors delayed publication of the API ratings by which teacher and school bonuses were determined. | Huntington Beach curriculum and instruction director Lynn Bogart responded, "I think there is so much data being required and the timeline is so short for all parties…I'm amazed that we're doing as well as we're doing" (Tully Tapia, 2000, p. 2). |
| (9) 2001 | MCAS/ Massachusetts DOE | Local school administrators | The DOE averaged the percentage of student failures over two years and used these to rate schools. Averaging percentages is not recommended in statistical computations because it yields inaccurate results. | The DOE claimed that the procedure altered school grades very little. Still, they agreed to recalculate the averages of school districts that requested it (Maffei, 2001). |
| (10) 2001 | UK Department on Education and Employment and Edexcel | UK Department for Education and Employment and Edexcel | A mistake made in calculating school averages in 1997 resulted in British primary schools being incorrectly rewarded in 2001. Estelle Morris, School Standards Minister for England's Department for Education and Employment, indicated that the miscalculation cost her department 2 million pounds in extra award money (Two million pounds, 2001). Primary schools in England received reward money and a certificate of improvement if they increased their score point average on two tests by a certain percentage. | The affected schools were allowed to keep the reward money (an average of £6,500 per school), but were notified that their names would be removed from the list of improvement winners and they would therefore not be sent certificates. |

[1] *California created a list of 100 "similar schools" to determine school rankings. Through this list, schools were categorized by socioeconomic data that included information about parents' education levels. Estimates of the educational levels were provided by students who filled out a questionnaire that accompanied the state exam. This method of ranking schools was not fully disclosed, as required by California's Brown Act (California First Amendment Coalition, 2001). Soon after the error was discovered, a Pasadena parent filed a lawsuit against the DOE in an attempt to gain access to the ranking data. The DOE finally released the records along with instructions on how to calculate the ratings; however, individuals were left to figure out on their own how the 100 "similar schools" were identified (Shafer, 2000a).*

# APPENDIX C:
## Errors in School Rankings

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (11) 2001 | FCAT/ Florida DOE | School principal | The DOE changed a school's grade from a C to an A after the building principal notified them that a high-scoring student's test score had been left out of the school average. After including the student's test in the school's score, the percentage of students passing changed from 49.5% to 50%. In a *Palm Beach* editorial that criticized the DOE's school grading system, it was noted that it was not unusual for a school's grades to fluctuate one or two grades from one year to the next (Editorial: FCAT's funny math, 2001). | |
| (12) 2001 | Colorado State Assessment Program (CSAP)/ Colorado DOE | Local education officials | Many data errors were found on Colorado's state school report cards. Among the mistakes reported were: statistics on students, teachers, and administrators were incorrect, test scores were inaccurate, and school officials believed that some of the state rankings were wrong. District officials were particularly concerned about the veracity of the rankings because the method for calculating them was undisclosed (Kreck, 2001; Hubler & Whaley, 2001). | Colorado school ratings had high stakes attached to them. Schools rated as "unsatisfactory" would be eligible for grants and pay incentives for three years, then subject to removal of staff in the fourth year if there was no improvement (Hubler, 2001). |
| (13) 2001 | Stanford 9 Achievement Tests (SAT-9)/ Harcourt Educational Measurement | Employees from a Fresno County school district | Questioned by local school officials, Harcourt admitted to using the wrong set of norms in calculating some California student scores and school averages – an error that inflated scores in 22 schools. In six of the schools, $750 thousand had been erroneously paid out in bonus money to both staff and schools. In the other 16, staff were informed that they were no longer eligible for bonuses, or that their bonuses would be reduced (Groves & Smith, 2001; Herendeen, 2001; Lopez, 2001; Scoring error sends cash, 2001). | Harcourt blamed the error on the tight exam schedule and said that in the future they would rather pay a penalty for providing late results than submit faulty scores (Groves & Smith, 2001). Educators who were asked to return the money were reportedly upset because many had already spent it (Lopez, 2001). A Harcourt official placed responsibility for the financial problems incurred by the error with the DOE as the bonus program was state-run (Herendeen, 2001). |
| (14) 2002 | TAAS/ Houston Independent School District (HISD) | Staff at HISD Vanguard middle school | A clerical error in which students were erroneously designated as dropouts caused teachers at a Houston middle school to miss out on $800 per-person bonus money. When the error, which was made by a worker at the school, was discovered, school employees petitioned HISD to issue the award (Markley, 2002). | |

# APPENDIX C:

## Errors in School Rankings

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (15) 2002 | FCAT/ Clay County School District | Superintendent of schools | Clay County superintendent of schools blamed the use of an old computer program for a data entry error that resulted in the state grades of ten district schools to be either incomplete or too low. Upon detection of the error, the superintendent petitioned the Florida DOE to reconsider the grades (Cravey, 2002). | |
| (16) 2002 | FCAT/ Florida DOE & NCS Pearson | Officials from affected school districts | Dozens of school districts received no state grades for their 2002 scores. Of the 124 schools that were ungraded, some were only open for one year and were ineligible for grading, and others received no grades because of a "programming error" at the DOE (Fischer, 2002). One Montessori school never received scores because NCS Pearson claimed not to have received the tests (Haller, 2002). | |
| (17) 2002 | Ohio Proficiency Tests (OPTs)/ Ohio DOE | Ohio DOE | The DOE mistakenly included 203 of 415 elementary schools on their list of low-performing schools. They blamed the problem on a computer programming error that required schools to submit scores showing increases in both fourth- and sixth-grade, even though hundreds of schools did not have both of these grades (Candisky, 2002). | The DOE designated schools as low-performing this year, in compliance with the federal government's *No Child Left Behind Act*. Parents of students in low performing schools were to be notified of this designation soon after the scores were released so that they could opt to enroll their children in other public schools (State identified 203 schools, 2002). |
| (18) 2002 | Colorado's School Ranking Program/ Colorado DOE | Colorado's DOE | Five school ratings were upgraded as a result of an omissions error in which the results of Spanish language tests were not included as part of the schools' ranking data. Four were changed from "unsatisfactory" to "low" and another moved from "low" to "average" (Mitchell, 2002). | The DOE warned that more ratings could be changed (up or down) as their staff examined the impact of the error. |

# APPENDIX D:
# Gray Areas; Test Results that Don't Add Up

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (1) 1993 | Norm-Referenced Assessment Program for Texas (NAPT)/ Riverside Publishing Company | Houston Independent school district officials | Houston officials observed large fluctuations in percentile scores on the NAPT exam. For example, in 1992 Houston 11th graders scored at the 34th national percentile; they scored at the 46th percentile after officials questioned the score and Riverside re-graded them. In 1993, Houston students scored at the 58th percentile, on average. A similar pattern of fluctuations was noted in a Houston elementary school's third-grade NAPT scores: in 1992, 18% of the students performed above grade level while 66% did so the following year – during which time the school's scores on another test (the TAAS) dropped sharply (Markley, 1993). | Before TAAS, Texas used NAPT to measure achievement. The legislature voted to expand the TAAS and phase out the NAPT in part, because district employees complained that test scores on the NAPT were unreliable. Frank Petruzielo, then superintendent of Houston said, "I assure you that if… the NAPT is voluntary, we won't be volunteering" (quoted in Markley, 1993, p. 3). |
| (2) 2000 | SAT-9, Form T/ Harcourt Brace | Officials at State DOEs in: CA, FL, AL, AZ, DE, and SD | From 1998-2000, ninth- and tenth-grade national percentile scores on the SAT-9 Form T reading tests were significantly lower than eighth-grade scores (Nguyen, 1998; Smith, 1998; Hegarty, 2000a; Hirschman, 2000). The drop was documented in six states that used the test: California, Florida, Alabama, Arizona, Delaware, and South Dakota (Hoff, 2000), yet officials in some states claimed that the drop in reading scores paralleled similar declines seen on other standardized tests. Officials at Harcourt Brace maintained that the test form was technically adequate (Hoff, 2000; Schrag, 2000). | Florida DOE officials were reluctant to release SAT-9 high school scores in the fall of 2000 because scores were so much lower than predicted. California DOE administrators requested that Harcourt Brace hire an independent evaluator to determine if the test was flawed (Hoff, 2000). To date, nothing has been done to alter the test and no explanation has been found (Hegarty, 2000b). |
| (3) 2000 | Elementary School Proficiency Assessment (ESPA)/ National Computer Systems | Officials at New Jersey's DOE | In 2000, scores on New Jersey's fourth-grade Language Arts Literacy test were dramatically lower than scores for other subject matter tests and lower than language arts scores in other grades. For general education students, the language mean score was 202.4, while the mathematics and science mean scores were 219.4 and 234.3, respectively (New Jersey Department of Education, 2001, p. 5). DOE officials singled out this test as having the most serious score discrepancies (Johnston, 2000). | New Jersey Education Commissioner, David Hespe, asked the contractor to investigate whether the test items were appropriate, and if so, whether the scoring was accurate. To date, no explanation has been found (Johnston, 2000). |

# APPENDIX D:

## Gray Areas; Test Results that Don't Add Up

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (4) 2000 | Maryland State Performance Assessment Program (MSPAP)/ Developed by Maryland DOE, Scored by Measurement Inc. | Maryland Department of Education | A discrepancy between the seven-year score gains on the eighth-grade reading test and other Maryland state assessments spurred a cottage industry of reading specialists in Maryland's middle schools. While the seven-year accumulated score gains for the other eighth-grade state assessments were nine percentage points or higher, gains on the reading test were less than one percentage point (Libit, 2000). | The Abell Foundation questioned the content validity of the reading tests, recommending that divergent, open-response questions be replaced with content-specific basic skills questions (such as those that address phonics or vocabulary acquisition). Maryland DOE officials countered that the results from the MSPAP were consistent with those of another nationally-standardized test, the Comprehensive Test of Basic Skillls (Bowler, 2000). In March 2002, the state superintendent announced that the administration of the eighth-grade exams would be "optional" (Libit, 2002). Additional problems plagued the program in 2002 when large numbers of school districts saw severe drops in their test scores (Keller, 2002). Following these declines, the DOE announced that they would stop the MSPAP and adopt a more traditional testing program (Hoff, 2002). |
| (5) 2000 | Oklahoma State Teachers Test/ National Evaluation Systems | Local school super-intendents | A possible error involving passing rates surfaced in Oklahoma on the state teacher's exams. A failure rate of about 30% on two of the exams left a shortage of teacher candidates. The state commission responsible for governing the teacher assessment program planned to alter some of the exams after meeting with the testing contractor. "Commission officials said the low scores [were] likely due to miscalculations in scoring rather than a poor pool of teachers. Oklahoma has been recognized recently for the high quality of its educators" (Plumberg, 2000, p. 1). | As of spring 2001, no alterations to the test had been indicated on the Teacher Candidate Assessment web page (Oklahoma Commission for Teacher Preparation, 2001). |

# APPENDIX D:

## Gray Areas; Test Results that Don't Add Up

| Year of Discovery | Test and Testing Contractor or Organization Responsible | Error Found by | Description of Error | Response to Error |
|---|---|---|---|---|
| (6) 2000 | Massachusetts Comprehensive Assessment System (MCAS)/ Massachusetts DOE/Harcourt Brace | | Results on the fourth-grade English language arts (ELA) portion of the MCAS were consistently low from 1998-2000. During this time, the level of fourth-graders scoring in the "failing" and "needs improvement" categories remained virtually unchanged at 80% (Massachusetts Department of Education, 2000). A report issued by the National Board on Educational Testing and Public Policy (Horn et al., 2000) found that, at one school, a significant number of students scoring above the 60th percentile on the fourth-grade Educational Records Bureau (ERB) reading exam (described as a difficult test) scored at the "needs improvement" level on MCAS. One student scoring at the 80th percentile on the ERB failed the fourth-grade ELA portion of MCAS. Also significant is that while only 20% of the fourth-graders scored "proficient" or above on the MCAS ELA exam in 2000, 62% of eighth-graders scored at this level during the same year (Massachusetts Department of Education, 2000). A report on the fourth-grade ELA test by Stokes and Stokes (2000) criticized the 1998 and 2000 exams for containing items that were extremely difficult for these students to respond to appropriately. | Chris Martes, director of the Massachusetts Association of School Superintendents, questioned the disparity between the fourth- and eighth-grade scores, "How do you reconcile the fact that in eighth grade a much higher number of the kids are in "advanced"? [MCAS] has been like that the three times they've administered it, and they've never fixed it" (quoted in Griffin, 2000). In a 2001 memo, the Massachusetts Commissioner of Education announced that the fourth-grade MCAS performance levels in reading would be reviewed (Driscoll, 2001). The proportion of students scoring at the "failing" and "needs improvement" levels was originally reported to be 80% in 2000. After the DOE adjusted the fourth-grade performance levels, the proportion of fourth-graders scoring in the bottom two performance categories dropped to 51% for that year (Massachusetts Department of Education, 2000 & 2002a). |
| (7) 2002 | Regents physics exam/New York Regents | Students and teachers | Pass rates on the Regents physics exam plummeted from 88% in 2001 to 67% in 2002 after the exam underwent a number of changes in the types of questions it asked and how students were scored. Superintendents, concerned that their students would be turned away from selective universities because of the low scores, sent out a letter to the institutions stating, "We believe the physics Regents exam grade to be suspect" (Cardinale, 2002, p. 1). Though the state gave school districts seven months to adjust to the changes, the exam was apparently administered without a thorough evaluation of the impact of the changes. | At the time of this report, ten school districts were suing the state over the low scores. This and other problems on the Regents resulted in a state review of the process for creating these tests as well as beefed-up training for teachers who develop test questions (Gormley, 2002). Students who did poorly were offered a retest (Hughes, 2002). |

# BIBLIOGRAPHY


AIMS writing test scores delayed. (2001, October 16). *The Associated Press State & Local Wire*. Retrieved June 26, 2002, from Lexis-Nexis database.

Allen, A. W. (1999, June 15). Student tests to be re-scored; Officials say writing grades too high. *The Burlington Free Press*. p.1.

Allen, Lamar, and Board of Trustees for Alabama State University v. The Alabama State Board of Education, No. 97-6808, (11th Cir. 1999).

Akin, P. (2002, August 14). SOL scorer erred; Writing scores better: 5,625 who 'failed' now are passing. *The Richmond Times-Dispatch*. Retrieved September 8, 2002, from Lexis-Nexis database.

Archibold, R. C. (1999, September 16). 8,600 in summer school by error, board says. *The New York Times*. Retrieved September 16, 1999, from: http://www.nytimes.com

Arizona Department of Education. (2000, March 17). *Some 1999 AIMS math scores to be revised*. Retrieved April 15, 2000, from: http://www.ade.state.az.us

Asimov, N. (1999a, July 12). Schools give test publisher failing grade; State exam results scrambled, delayed. *The San Francisco Chronicle*. Retrieved April 27, 2001, from Lexis-Nexis database.

Asimov, N. (1999b, October 12). School testing firm face increased scrutiny; Davis signs law to make scoring more efficient. *San Francisco Chronicle*. Retrieved October 12, 1999, from: http://www.sfgate.com./chronicle

Bach, L. K. (1999a, September 18). Testing company botches report. *Las Vegas Review-Journal*. Retrieved September 20, 1999, from: http://www.lvrj.com

Bach, L. K. (1999b, October 20). Errors wrong schools: Three institutions come off the state's danger list after botched results from a 1999 test are corrected. *Las Vegas Review-Journal*. Retrieved October 27, 1999, from: http://www.lvrj.com

Bakst, B. (2000, October 12). Months late, graduates get a handshake from the governor. *Associated Press*. Retrieved May 24, 2002, from Lexis-Nexis database.

Barbour, I. (1993). *Ethics in an age of technology*. The Clifford Lectures: Volume Two. San Francisco: Harper.

Beaton, A., Zwick, R., Yamamoto, K., Mislevy, R., Johnson, E., and Rust, K. (1990). *The effect of changes in the National Assessment: Disentangling the NAEP 1985-1986 reading anomaly* (No. 17-TR-21). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Begley, S., and Carey, J. (1981, April 6). A sunshine law for SAT's. *Newsweek*. Retrieved May 23, 2001, from Lexis-Nexis database.

Benning, V., and Mathews, J. (2000, September 8). Statewide scores up on most Va. tests; Higher standards are working, governor says. *The Washington Post*. Retrieved November 13, 2001, from: http://www.washingtonpost.com

Bickman, L. (Ed.). (2000). *Research Design*. Thousand Oaks, CA: Sage Publications, Inc.

Bloom, B. S., Madaus, G. F., and Hastings, J. T. (1981). *Evaluation to improve learning*. New York: McGraw Hill Company.

Botched test handling renders EQAO useless; Education: Give Queen's Park an F-. (2001, December 6). *The Hamilton Spectator*. Retrieved November 4, 2002, from Lexis-Nexis database.

Bower, C. (1998, September 11). Error changes results of Missouri math test; Analysis shows fewer students with high scores. *St. Louis Post-Dispatch*. Retrieved February 23, 2000, from Newslibrary database.

Bowler, M. (2000, November 19). MSPAP is imperfect but gets passing grade. *Sunspot.net*. Retrieved November 29, 2000, from: http://sunspot.net

Bowman, D. H. (2000, September 6), Minn. extends testing contract despite scoring mistakes. *Education Week*. Retrieved September 5, 2000, from: http://www.edweek.org

Brandon, K. (1999, July 1). Reporting error inflates non-bilingual gains on tests. *San Diego Tribune*, Section 1, p. 8.

Bright, M., and McVeigh, T. (2002, September 1). Teachers fear A-level grades were 'fixed'. Concern grows that bright pupils may have lost university places. *The Observer*. Retrieved November 3, 2002, from: http://www.observer.co.uk

Brown, C. (1980, January 19). 163 students face repeat of SATs as scores vanish; Students repeating SATs after 1st scores vanish. *The Washington Post*. Retrieved April 27, 2001, from Lexis-Nexis database.

Brunts, J. (1999, January 14). Error in ISTEP+ scoring delays sending of results. *The Fort Wayne News-Sentinel*. Retrieved February 15, 2000, from Newslibrary database.

Buchanan, M. (2000, August 22). VUSD scores may change. *North County Times*. Retrieved September 5, 2000, from: http://www.nctimes.com

California First Amendment Coalition. (2001). *The Ralph M. Brown Act*. Retrieved April 27, 2001, from: http://www.cfac.org

Calvano, R. (1989, March 30). Schools due corrected scores on 1987 writing-skills test. *The San Diego Union-Tribune*. Retrieved April 27, 2001, from Lexis-Nexis database.

Campanile, C. (2001, July 9). 6th-graders got boost from testing blunder. *New York Post*. Retrieved June 11, 2001, from: http://www.nypostonline.com

Campbell, J. R., Hambo, C. M., and Mazzeo, J. (2000, August). *NAEP 1999 trends in academic progress: Three decades of student performance*. Retrieved January 5, 2003, from: http://nces.ed.gov/nationsreportcard//pubs/main1999/2000469.asp

Candisky, C. (2002, July 30). 'Bad schools' list drops by half; State goofed big time first time around. *The Columbus Dispatch*. Retrieved September 8, 2002, from Lexis-Nexis database.

Cardinale, A. (2002, November 8). Colleges are asked to ignore poor scores. *The Buffalo News*. Retrieved December 5, 2002, from Lexis-Nexis database.

Carlson, J. (2000, July 31). Hearing to look at skills-test scoring error; A key legislator will convene a meeting today to examine the misstep that kept more than 300 seniors from graduating. *Star Tribune*. Retrieved April 27, 2001, from Lexis-Nexis database.

Cassidy, S. (2000, April 7). Exam boards told to improve service. *The Times Educational Supplement*. Retrieved December 4, 2002, from Lexis-Nexis database.

Children, Families and Learning. (2000a, July 28). *Commissioner Jax blasts NCS for scoring errors on February and April 2000 math tests; Demands fixes at NCS expense*. Retrieved August 14, 2000, from: http://www.educ.state.mn.us

Children, Families and Learning. (2000b, July 28). *Remarks of Commissioner Christine Jax, Ph.D.* Retrieved August 14, 2000, from: http://www.educ.state.mn.us

Chrismer, E. (1999, July 23). Scores up on latest reading, math test English-speaking students in state perform above national average. *Modesto Bee*. Retrieved January 18, 2003, from Lexis-Nexis database.

Clare, J. (2000, April 4). Must try harder, exam boards told. *Electronic Telegraph*. Retrieved April 30, 2000, from: http://www.telegraph.co.uk

Clarke, A. (2000, August 12). Exams chief resigns. *BBC News*. Retrieved August 31, 2000, from: http://news.bbc.co.uk

Coleman, S. (2000, May 25). Printing error cited in latest round of MCAS tests. *The Boston Globe*. Retrieved May 27, 2000, from: http://www.boston.com./dailyglobe

Collins, R. (2001, October 4). MCAS changes to affect high, low scores, 4th graders, and comparisons. *TownOnline.com*. Retrieved November 13, 2001, from: http://www.townonline.com

Colvin, R. L. (1999a, July 2). Lack of oversight blamed for Stanford-9 test snafu. Education: Friction among state agencies and lack of clear authority over testing firm are among causes, officials say. *Los Angeles Times.com*. Retrieved November 5, 1999, from: http://www.latimes.com

Colvin, R. L. (1999b, July 7). New flaw is found in state test results. Education: Rankings for Long Beach's year-round schools compared with national average are wrong, publisher says. Other districts are urged to check data. *Los Angeles Times*. p. A3, A8.

Colvin, R. L. (2000, April 22). Rankings for many schools recalculated. Education: Many campuses move more than two rungs on 10-point performance scale after finding errors in reporting. *Los Angeles Times*. Retrieved April 30, 2000, from: http://www.latimes.com

Colvin, R. L., and Smith, D. (1999, June 30). Error may have skewed statewide test results: Education: Firm admits misclassifying 300,000 test-takers' English skills. L.A. scores are affected. *Los Angeles Times*. p. A1, A19.

Colvin, R., and Groves, M. (1999, July 1). State's students gain modestly in reading; Education: Partial Stanford 9 results show stronger improvement in math. Scoring blunder delays release of full data. *Los Angeles Times*. Retrieved May 16, 2002, from Lexis-Nexis database.

Corporate greed behind Minnesota test error. (Summer, 2002). FairTest Examiner.

Cravey, B. R. (2002, June 19). Schools challenge state grades; District error thought to have hurt. *The Florida Times-Union*. Retrieved June 20, 2002, from Lexis-Nexis database.

Crew backs company that erred on scores. (1999, December 1). *The New York Times*. Retrieved April 27, 2001, from Lexis-Nexis database.

Curran, J., and Drew, R. (1997, February 7). Student catches error in College Board test. *Chicago Sun-Times*. Retrieved April 27, 2001, from Lexis-Nexis database.

De Valle, E. (2001, May 31). Test scores of seniors thrown out. *The Miami Herald*. Retrieved June 1, 2001, from: http://www.miami.com./herald

de Vise, D. (1998a, August 5). Glitch lowers student FCAT scores; A computer error gave thousands of students statewide lower scores than earned on the important exam. *Tallahassee Democrat*. Retrieved February 23, 2000, from Newslibrary database.

de Vise, D. (1998b, August 5). Computer error lowers students' scores on test. *The Miami Herald*. Retrieved February 23, 2000, from Newslibrary database.

Dickie, K. (1987, July 13). Math teacher 'flunks' grammar questions. *Memphis Commercial-Appeal*.

Donsky, P. (2002, July 19). Student test scoring glitch throws schools into tizzy. *The Atlanta Journal Constitution*. Retrieved September 4, 2002, from Lexis-Nexis database.

Draper, N. (2000, August 16). Test denied diplomas for 54 at most; Scoring mess affected fewer seniors than first thought. *Star Tribune*. Retrieved April 27, 2001, from Lexis-Nexis database.

Draper, N. (2001, February 6). Error found in state reading test; The testing company repeated a letter in a list of answers; a state employee caught it. *Star Tribune*. Retrieved April 27, 2001, from Lexis-Nexis database.

Drew, D. P., Smetanka, M. J., and Shah, A. (2000, August 1). Blame spreads in test mess; Education officials admit response to parent's query took 2 months. *Star Tribune*. Retrieved April 27, 2001, from Lexis-Nexis database.

Drew, D. P., and Draper, N. (2001, April 15). State: No more test mess; Results of basic-skills reading, math exams due Wednesday. *Star Tribune* (Minneapolis). Retrieved April 27, 2001, from Lexis-Nexis database.

Driscoll, D. (2001, May 31). MCAS standard-setting recruitment for grade 3 reading, grades 4 and 7 English language arts, grade 6 mathematics. *Massachusetts Department of Education.* Retrieved June 16, 2001, from: http://www.doe.mass.edu/mcas/2001docs/0604memo.html

Edelman, S., and Graham, J. (1999, September 20). Test-scores debacle also hits Indiana schools. *New York Post*. Retrieved September 20, 1999, from: http://www.nypost.com

Editorial: FCAT's funny math. (2001, August 9). *PalmBeachPost.com*. Retrieved August 20, 2001, from: http://www.gopbi.com

The Educational Institute of Scotland (2000, October 13). The union's evidence. *The Times Educational Supplement, Scotland Plus*. p. 7.

Eng, L. (1991, June 14). Printing error puts students to the test – for second time. Education: Nearly 1,000 scholastic aptitude test scores in California were invalidated because the exam included some duplicated questions. *Los Angeles Times.* Retrieved April 27, 2001, from Lexis-Nexis database.

Error will force 1,000 to take SAT again. (1991, June 14). *Los Angeles Times*. Retrieved April 27, 2001, from Lexis-Nexis database.

Exam board admits 'printing error.' (2002, May 29). *BBC News*. Retrieved June 19, 2002, from: http://news.bbc.co.uk

Examiner sacked over GCSE row. (2001, October 25). *BBC News*. Retrieved November 5, 2002, from: http://news.bbc.co.uk

Fischer, K. (2000, September 26). Two schools earn A's after appeal. *St. Petersburg Times*. Retrieved September 26, 2000, from: http://www.sptimes.com

Fischer, K. (2002, June 14). 3 schools' report cards will be late from state. *St. Petersburg Times*. Retrieved June 26, 2002, from Lexis-Nexis database.

Fiske, E. (1981a, March 27). "Truth in Testing" to be nationwide. *The New York Times.* Retrieved April 27, 2001, from Lexis-Nexis database.

Fiske, E. (1981b, April 14). Pyramids of test question 44 open a Pandora's Box. *The New York Times*. Retrieved April 27, 2001, from Lexis-Nexis database.

Fiske, E. (1981c, March 24). A second student wins challenge on answer to math exam problem. *The New York Times.* Retrieved April 27, 2001, from Lexis-Nexis database.

Fiske, E. (1981d, March 17). Youth outwits merit exam, raising 240,000 scores. *The New York Times*. Retrieved April 27, 2001, from Lexis-Nexis database.

Flannery, P. (2000, October 8). AIMS: Story of missteps, false starts. *The Arizona Republic*. Retrieved October 12, 2000, from: http://www.arizonarepublic.com

Fletcher, M. A. (2001, July 9). As stakes rise, school groups put exams to the test. *The Washington Post*. Retrieved July 9, 2001, from: http://www.washingtonpost.com

Florida Department of Education. (2001). *State board of education administrative rules.* Retrieved June 16, 2001, from: http://www.firn.edu

Florida Department of Education. (2002, March 27). *Technical report: For operation test administrations of the 2000 Florida Comprehensive Assessment Test.* Retrieved December 3, 2002, from: http://www.firn.edu/doe/sas/fcat/fcatpub2.htm

Ford, R. (1997, February 7). Smarter than the SAT; N.H. high schooler discovers a mistake. *The Boston Globe.* p. A1, A21.

Foucault, M. (1979). *Discipline and punish: The birth of the prison.* (Alan Sheridan, Trans.). News York: Vintage.

Frahm, R. (1992, February 3). Performance testing backed despite mixup. *The Hartford Courant.* Retrieved April 27, 2001, from Lexis-Nexis database.

Frahm, R. (1999, July 28). Glitch distorts scores on SATS; Hundreds of test-takers get an unexpected boost. *The Hartford Courant*. Retrieved February 23, 2000, from Newslibrary database.

Franck, M., and Hacker, H. (2000, March 7). Ladue schools' challenge of testing shows pitfalls in the system; Some critics call for overhauling the way tests are processed. *St. Louis Post-Dispatch*. Retrieved April 27, 2001, from Lexis-Nexis database.

Gallagher, E. D. (2001). *A compilation of problems in 10th grade MCAS math tests.* Retrieved October 10, 2001, from http://www.es.umb.edu/edgwebp.htm

Gehring, J. (2001, October 24). Mass. school policies praised as test scores rise. *Education Week.* Retrieved November 13, 2001, from: http://edweek.org

Gendar, A. (1999, September 22). Testing firm's snafu cost the city $3.8M. *Daily News* (New York). Retrieved April 27, 2001, from Lexis-Nexis database.

Gendar, A. (2002, June 27). Reading test scores held back. *Daily News* (New York). Retrieved September 12, 2002, from Lexis-Nexis database.

Gewertz, C. (2000, November 1). N.M. flubs list of most-improved schools. *Education Week.* Retrieved November 2, 2000, from: http://www.edweek.org

Gewertz, C. (2001, June 20). Test firm, N.Y.C. officials say scores were 'overstated.' *Education Week*. Retrieved June 19, 2001, from: http://www.edweek.org

Glass, G. (1977). Standards and criteria. *Occasional Paper Series*, 10. Retrieved November 5, 2002, from: http://www.wmich.edu/evalctr/pubs/ops/ops10.html

Gledhill, L. (1999, August 3). State Board of Education fines testing company $1.1 million; Confidence lost in scores after run of mistakes. *The San Francisco Chronicle*. Retrieved April 27, 2000, from Lexis-Nexis database.

Good, J. (1998, October 9). Errors delay state test-score data; 'Cut-and-paste' mistake blamed. *The Burlington Free Press*. p. 1.

Goodbye GCES? (2001, June 5). BBC News. Retrieved June 26, 2002, from: http://news.bbc.co.uk

Goodnough, A. (2001, June 9). School board in dispute on test scores. *The New York Times.* Retrieved October 8, 2002, from Lexis-Nexis database.

Goodnough, A. (2002, September 28). After disputes on scoring, school system switches provider of reading tests. *The New York Times*. Retrieved October 1, 2002, from: http://www.nytimes.com

Gormley, M. (2002, August 16). AP Interview: Chancellor says Regents exams put to test. *The Associated Press State & Local Wire*. Retrieved September 12, 2002, from Lexis-Nexis database.

Grant, A. H. (2002, September 21). Students can seek punitive damages from testing firm. *Pioneer Press* (St. Paul). Retrieved September 23, 2002, from: http://www.twincities.com

Greenberger, S. (2001, October 25). Mixed signals from latest MCAS scores. *The Boston Globe*. Retrieved November 13, 2001, from: http://www.boston.com

Greenberger, S., and Dedman, B. (2001, November 2). Schools make big gains on MCAS. *The Boston Globe*. Retrieved November 13, 2001, from: http://www.boston.com

Griffin, R. (2000, December 4). Teachers: 4th-grade test too tough. *Eagle-Tribune*. Retrieved December 5, 2000, from: http://www. Eagletribune.com

Groves, M. (1999, July 2). 9th-grade dip on tests baffles educators. *Los Angeles Times,* p. A3, A26.

Groves, M. (2000, September 13). California and the west; State rewards for schools are delayed. Education: Flawed data on participation rates on Stanford 9 tests proves a hurdle for achievement bonus payouts. *Los Angeles Times.* Retrieved April 27, 2000, from Lexis-Nexis database.

Groves, M., and Smith, D. (2001, September 28). Ineligible schools got rewards. Test: Scoring foul-up sent $750,000 to the wrong campuses. Others will get smaller amounts. *Los Angeles Times*. Retrieved October 7, 2001, from: http://www. latimes.com

Grow, D. (2000, August 2). Bungling bureaucrats put a spin on test fiasco; Now that more students have passed, state officials working to pass the buck. *Star Tribune*. Retrieved April 27, 2000, from Lexis-Nexis database.

Gruss, M. (1999, November 14). Can ISTEP be saved? Publishers' errors ruin credibility. *The Journal Gazette*. Retrieved February 15, 2000, from Newslibrary database.

Haller, G. (2002, July 6). Montessori FCAT tests disappear. Keysnews.com. Retrieved August 5, 2002, from CARE listserve: care@yahoogroups.com

Haney, W., Fowler, C., Wheelock, A., Bebell, D., and Malec, N. (1999, February 11). Less truth than error? An independent study of the Massachusetts Teacher Tests. *Education Policy Analysis Archives, 7*;4. Retrieved February 20, 2003, from: http://epaa.asu.edu/epaa/v7n4/

Haney, W. (2000, August 19). The myth of the Texas miracle in education. *Education Policy Analysis Archives*, *8*;41. Retrieved February 20, 2003, from: http://epaa.asu.edu/epaa/v8n41/

Harasta, C. (2000, September 21). Cathy Harasta's Olympic sports column: Equipment error doesn't measure up to standards. *The Dallas Morning News*. Retrieved July 2, 2001, from: http://www.dallasnews.com

Harp, L. (1997a, July 22). Test snafu deepens: Scoring formula missing. L*exington Herald-Leader*. Retrieved February 23, 2000, from the Newslibrary database.

Harp, L. (1997b, September 11). KIRIS recount lifts 4th-grade scores but changes little. *Lexington Herald-Leader*. Retrieved October 17, 1999, from Newslibrary database.

Harris, S., and Clark, L. (2002, October 16). A-levels: No one to blame (again); Out of 91,000 pupils, only 168 lost their place at first college. After a month of controversy and confusion which left students stranded and tarnished the 'gold standard' exam, the usual bout of buck-passing. *Daily Mail* (London). Retrieved November 3, 2002, from Lexis-Nexis database.

Harrison, A. (2002, September 20). A-level report blames teaching. *BBC News.* Retrieved November 3, 2002, from: http://news.bbc.co.uk

Hartocollis, A. (1999a, January 18). New reading test emerges battered, but triumphant. *The New York Times.* Retrieved September 27, 1999, from: http://www.nytimes.com

Hartocollis, A. (1999b, May 26). Most of state's 4th graders fail new English test. *The New York Times*. Retrieved September 27, 1999, from: http://www.nytimes.com

Hartocollis, A. (1999c, September 15). Chancellor cites score errors. *The New York Times*. Retrieved September 16, 1999, from: http://www.nytimes.com

Hartocollis, A. (1999d, September 17). Miscalculation on scores shows a weakness of tests. *The New York Times*. Retrieved September 20, 1999, from: http://www.nytimes.com

Hartocollis, A. (2001, June 30). Math test needed for high school graduation had confusing errors, state officials say. *The New York Times.* Retrieved November 7, 2001, from Lexis-Nexis database.

Hayes, D. (2002, September 19). Unions welcome inquiry move. *The Press Association Limited.* Retrieved November 3, 2002, from Lexis-Nexis database.

Hayes, D., and Linden, M. (2002, October 2). Exam grades review for thousands of students. *The Press Association Limited*. Retrieved November 3, 2002, from Lexis-Nexis database.

Hayward, E. (2001a, May 23). Test creator contradicts policy. Harcourt Chief: MCAS should not be sole requirement for graduation. *The Boston Herald*. Retrieved May 23, 2001, from: http//www.bostonherald.com

Hayward, E. (2001b, October 4). Some MCAS results to be readjusted due to glitch. *The Boston Herald*. Retrieved November 13, 2001, from: http://www.bostonherald.com

Heads should roll over this: Gov. Wilson must personally probe state education department for test scandal. (1994, April 12). *Los Angeles Times*. Retrieved April 27, 2001, from Lexis-Nexis database.

Hegarty, S. (2000a, June 30). Officials withhold test scores; The Inspector General's Office will investigate why Florida reading scores were so inconsistent. *St. Petersburg Times*. Retrieved June 30, 2000, from: http://www.sptimes.com

Hegarty, S. (2000b, September 20). Why low test scores? Who knows? *St. Petersburg Times*. Retrieved March 27, 2000, from: http://www.sptimes.com

Helderman, R. S., and Keating, D. (2002, November 8). Two-thirds of schools meet Va. standards. In the region, every district improved and most reached goal for full accreditation, SOL shows. *The Washington Post*. Retrieved December 28, 2002, from Lexis-Nexis database.

Henderson, D., and Munro, N. (2000, October 13). Higher post mortem makes sorry reading. *The Times Educational Supplement, Scotland*. No. 1770, p. 4-5.

Hendrie, C., and Hurst, M. (2002, September 4). Error on tests in Nevada and Georgia cost publisher Harcourt. *Education Week*. Retrieved September 6, 2002, from: http://www.edweek.org

Herendeen, S. (2001, October 3). Teachers may lose expected bonuses. Firm's mistake gave 10 local schools inflated Stanford 9 test results. *Modesto Bee* (California). Retrieved June 26, 2002, from Lexis-Nexis database.

Henriques, D. (2001, May 25). Testing service error lowered some scores on entrance exam. *The New York Times*. Retrieved May 31, 2001, from: http://www.nytimes.com

Hirschman, B. (2000, September 19). State officials still stumped by dip in FCAT reading scores for 9th, 10th graders. *Sun-Sentinel*. Retrieved September 20, 2000, from: http://www.sun-sentinel.com

Hoff, D. J. (2000, November 8). States question national reading-test scores. *Education Week*. Retrieved November 10, 2000, from: http://www.edweek.org

Hoff, D.J. (2002, April 3). Md. To phase out innovative testing program. *Education Week*. Retrieved May 20, 2002, from: http://www.edweek.org

Horn, C., Ramos, M., Blumer, I., and Madaus, G. (2000, April). Cut scores: Results may vary. *National Board on Educational Testing and Public Policy Monographs*, *1*;(1).

Houtz, Jolayne. (1999a, August 5). Scoring flub mars results of state academic test. *Seattle Times*. Retrieved December 5, 1999, from: http://www.seattletimes.com

Houtz, J. (1999b, September 1). Error on student essays to cost firm $600,000. *Seattle Times*. Retrieved September 1, 1999, from: http://www.seattletimes.com

Hubler, E. (2001, September 13). School rankings released. *DenverPost.com*. Retrieved September 18, 2001, from: http://www.denverpost.com

Hubler, E., and Whaley, M. (2001, September 25). Report card errors stun schools. *DenverPost.com*. Retrieved September 25, 2001, from: http://ww.denverpost.com

Hughes, C. (2002, August 14). Physics students take another shot; State offers them a second chance at retooled test after many did poorly in June. *The Times Union* (Albany, NY). Retrieved September 12, 2002, from Lexis-Nexis database.

Institute of Medicine. (1999, November 29). Preventing death and injury from medical errors requires dramatic, system-wide changes. *The Medical Reporter*. Retrieved January 30, 2003, from: http://medicalreporter.health.org/tmr1199/medical_errors.html

Institute of Medicine. (2000). *To err is human: Building a safer health system*. Washington, D. C.: National Academy Press.

ISTEP+ faces an increasing wave of critics. (1999, September 20). *Indiana Messenger-Inquirer.com.* Retrieved November 5, 1999, from: http://www.messenger-inquirer.com

Jackson, P. (1999, October 20). Wrong chart caused student test-scores recall. Employee used data from 1998 to determine grades for 1999. *The News Journal* (Delaware).

Johnston, Robert C. (2000, September 6). New Jersey investigating suspiciously low 4th-grade test scores. *Education Week*. Retrieved September 8, 2000, from: http://www.edweek.org

Jones, R., and Mezzacappa, D. (1998, January 13). Testing slighted students' abilities. The revelation renewed criticism of the process of rewards and sanctions based on test results. *Philadelphia Inquirer*. Retrieved January 13, 1998, from: http://www.phillynews.com

Kane, T., and Staiger, D. (2001, March). *Improving school accountability measures* (Working Paper 8156). Cambridge, MA: National Bureau of Economic Research.

Keller, B. (2002, February 6). Controversy surrounds release of Maryland test results. *Education Week.* Retrieved May 20, 2002, from: http://www.edweek.org

Kiernan, L. (1979, June 30). Do not fold, staple…; Computer 'crinkles' bar score sheet, erroneously fails Baltimore graduate. *The Washington Post*. Retrieved April 27, 2001, from Lexis-Nexis database.

King, S. (1999, January 13). Mistakes found in scores of ISTEP; 3 Allen County school districts reported errors. *The Fort Wayne News-Sentinel*. Retrieved February 15, 2000, from Newslibrary database.

King, K., and White, S. E. (2002, August 14). 1,900 students' scores on writing test upgraded. *The Virginian-Pilot*. Retrieved September 8, 2002, from Lexis-Nexis database.

Klampe, M. L. (1999, September 16). Indiana finds score error in comparisons of ISTEP+; Students' results vs. the norm were miscalculated. *The Fort Wayne News-Sentinel.* Retrieved February 15, 2000, from Newslibrary database.

Klampe, M. L. (2000, February 12). ISTEP scoring error relief for some; The discovery means 13 students actually passed the exam. *The Fort Wayne News-Sentinel.* Retrieved January 15, 2000, from Newslibrary database.

Kossan. P. (2000a, November 22). Keegan backs off AIMS requirements. *The Arizona Republic*. Retrieved November 22, 2000, from: http://www.arizonarepublic.com

Kossan P. (2000b, November 27). By trying too much too quick, AIMS missed mark. *The Arizona Republic*. Retrieved November 27, 2000, from: http://www.arizonarepublic.com

Kossan, P. (2002, June 4). McGraw Hill wants $263,000 from state. *The Arizona Republic*. Retrieved June 19, 2002, from: http://www.arizonarepublic.com

Kowal, J. (2001, June 9). Another error in Board of Ed. tests. Sixth-grade reading exams scored too high. *Newsday* (New York). Retrieved October 8, 2002, from Lexis-Nexis database.

Kreck, C. (2001, September 16). Littleton asking state to correct report cards. Documents will go out as printed, ed chief says. *DenverPost.com*. Retrieved September 18, 2001, from: http://www.denverpost.com

Kurtz, M., and Vaishnav, A. (2002, December 5). Student's MCAS answer means 449 others pass. *Boston.com*. Retrieved December 5, 2002, from: http://www.boston.com

Kurvers, et. al. v. National Computer Systems, Inc., MC 00-11010 (Hennepin, MN. 2002).

Lake, R. (2002, July 30). Scoring mistake affects 736 students. *Las Vegas Review-Journal*. Retrieved September 6, 2002, from Lexis-Nexis database.

Larkin, N. (2000, November 29). Maths exam gets its ABC (and D) wrong. *The New Zealand Herald Online*. Retrieved December 5, 2000, from: http://www.nzherald.co

Lawrence, D. (1999, July 14). Is medical care obsolete? *Kaiser Permanente*. Retrieved January 30, 2003, from: http://www.kaiserpermanente.org./newsroom/releases/speech.html

Levinson, A. (2001, July 8). Major teacher-testing company won't release data. *Sentinel and Enterprise* (Leominster, MA). A1, 4.

Libit, H. (2000, November 26). Reading scores vex educators; 8th-graders perform consistently poorly in MSPAP section; Seven-year trend; Officials intensify efforts to boost pupils' abilities. *Sunspot.net*. Retrieved November 27, 2000, from: http://www.sunspot.net

Libit, H. (2002, March 5). Districts will get choice on MSPAP. State won't require schools to test 8th-graders in spring. *The Baltimore Sun*. Retrieved May 20, 2002, from Lexis-Nexis database.

Lindsay, J. (2001, May 23). Students find mistake in math portion of MCAS test. *Associated Press*. Retrieved May 23, 2001, from: http://www.masslive.com

Lines, D. (2000, April 7). A disaster waiting to happen. [Op-ed]. *The Times Educational Supplement*. Retrieved December 4, 2002, from Lexis-Nexis database.

Lopez, P. (2001, September 29). Schools angry at possible refund. Districts that got money for test scores may have to return it. *The Fresno Bee* (California). Retrieved June 26, 2002, from Lexis-Nexis database.

Lu, A. (2002, May 23). Foul-ups put online exams to the test; End-of-grade series off to a shaky start. *Newsobserver.com*. (North Carolina). Retrieved May 26, 2002, from: http://newsobserver.com

Ludlow, L. (2001). Teacher test accountability: From Alabama to Massachusetts. *Education Policy Analysis Archives, 9;*6. Retrieved February 22, 2001, from: http://epaa.asu.edu/epaa/v9n6.html

Lyall, S. (2001, December 19). More deaths in England due to error, report says. *The New York Times*. Retrieved May 21, 2002, from: http://www.nytimes.com

MacBride, G. (2000). Beyond the 2000 exams disaster. *Scottish Education Journal, 84*;5, p. 6-7.

Macdonald, K. (2000a, August 9). From troubled beginnings. *BBC News*. Retrieved August 31, 2000, from: http://news.bbc.co.uk

Macdonald, K. (2000b, August 18). SQA puts figure on exam mistakes. *BBC News*. Retrieved August 31, 2000, from: http://news.bbc.co.uk

Macdonald, K. (2000c, August 31). Exam checks not carried out. *BBC News*. Retrieved August 31, 2000, from: http://www.news6.thdo.bbc.co.uk/hi/english/uk/scotland/newsid%5F9030833.stm

Mackay, A. (2000, August 23). Minister's new exams pledge. *BBC News*. Retrieved August 24, 2000, from: http://news.bbc.co.uk/hi/english/uk/scotland/newsid_891000/891933.stm

Mackenzie, A. (2000, September 2). Doubts cast over exam appeal deadline. *BBC News*. Retrieved August 31, 2000, from: http://news.bbc.co.uk

Maffei, G. (2001, January 26). Error found in scoring MCAS failure rates. *The Patriot Ledger* (South of Boston). Retrieved January 26, 2001, from: http://www.ledger.southofboston.com

Magee, M. (2002, April 19). Elephant essay test stirs confusion. Instruction flawed, school officials say. *The San Diego Union-Tribune*. Retrieved June 26, 2002, from Lexis-Nexis database.

Margin of error (2000, March 10). *St. Louis Post-Dispatch*. Retrieved April 27, 2001, from Lexis-Nexis database.

Markley, M. (1993, August 31). HISD suspects "congenital defects" in standardized testing. *The Houston Chronicle*. Retrieved April 27, 2001, from Lexis-Nexis database.

Markley, M. (2002, February 17). Clerical error causes T. H. Rogers teachers to miss bonus. HISD spokeswoman says district reviewing case. *Houston Chronicle*. Retrieved March 4, 2002, from: http://www.houstonchronicle.com

Marr, L. G., and Walters, J. (2001, December 5). 2,000 tests missing; KISS and tell book in the stores. D14; Disappearance of Grade 6 answer sheets latest foulup in provincewide school testing. *The Hamilton Spectator*. Retrieved November 4, 2002, from Lexis-Nexis database.

Massachusetts Department of Education. (2000, November). *Spring 2000 MCAS tests: Report of state results.* Retrieved December 5, 2000, from: http://www.doe.mass.edu/mcas/2000/results/re_ss.pdf

Massachusetts Department of Education. (2001, October). *Spring 2001 MCAS tests: State results by race/ethnicity and student status.* Retrieved November 5, 2002, from: http://www.doe.mass.edu/mcas/2001/results/re_ss.pdf

Massachusetts Department of Education. (2002a, August). *Spring 2002 MCAS tests: Summary of state results.* Retrieved November 5, 2002, from: http://www.doe.mass.edu/mcas/2002/results/summary

Massachusetts Department of Education. (2002b, December 4). *Points awarded to students who selected alternate answer on grade 10 MCAS exam*. Retrieved December 4, 2002, from: http://www.doe.mass.edu/news/news.asp?id=1328

Mathews, J. (2000a, May 30). Testing the Market. *The Washington Post*. Retrieved June 2, 2000, from: http://www.washingtonpost.com

Mathews, J. (2000b, June 4). SOL frenzy grips Va. campuses. *The Washington Post*. Retrieved November 13, 2001, from: http://www.washingtonpost.com

Measuring mix-up taints women's gymnastic event. (2000, September 20). *The New York Times*. Retrieved July 2, 2001, from: http://www.nytimes.com

Melendez, M. (2000, August 17). Error in figuring data cut school's rating, Texas official says. *Star-Telegram* (Fort Worth). Retrieved August 20, 2000, from: http://www.startext.net

Mendoza, M. (1999, September 18). CTB testing company faulted for inaccurate student scores. *Lexington Herald-Leader*. Retrieved September 20, 1999, from: http://www.kentuckyconnect.com

Merl, J. (1994, August 3). Problems undercut CLAS tests, panel says. Education: Statistics experts say sampling and scoring procedures need to be improved, and recommend that an outside firm be hired to administer the exams, but they praise program for innovation. *Los Angeles Times*. Retrieved April 27, 2001, from Lexis-Nexis database.

Meyer, L., Orlofsky, G. F., Skinner, R. A., and Spicer, S. (2002, January 10). The state of the states. *Education Week*. Retrieved November 6, 2002, from: http://www.edweek.org

Miles, T. (2002, September 20). Massive; Up to 70 A-level students affected at many schools. *The Evening Standard* (London). Retrieved November 3, 2002, from Lexis-Nexis database.

Mitchell, N. (2002, December 7). Error labeled four schools as failing. Spanish-language tests were thrown out by computer. *Rocky Mountain News*. Retrieved December 7, 2002, from: http://www.rockymountainnews.com

Mitchell, K. J., Robinson, D. Z., Plake, B. S., and Knowles, K. T. (Eds.). (2001). *Testing teacher candidates: The role of licensure tests in improving teacher quality*. Retrieved January 15, 2003, from: http://books.nap.edu/books/0309074207/html/R1.html/R1.html#pagetop

Mora, J. K. (1999, July 25). What do 'corrected' test scores reveal? *The San Diego Union-Tribune.* Retrieved April 27, 2001, from Lexis-Nexis database.

Moran, C. (1999, July 13). More scoring errors delay posting of students' test results. *The San Diego Union-Tribune.* Retrieved April 27, 2001, from Lexis-Nexis database.

Moran, C. (2000, July 23). School test gains may be illusory, critics say; Stanford 9 exams stays same, so familiarity can play role. *The San Diego Union-Tribune.* Retrieved April 27, 2001, from Lexis-Nexis database.

Morrison, J. (2000, November 30). Ministers escape exam farce blame. *BBC News.* Retrieved December 1, 2000, from: http://www.bbc.co.uk

Munro, N. (2000, October 13). SQA hunt for markers. *The Times Educational Supplement, Scotland.* No. 1770, p. 1.

Murphy, C. (1999, October 19). Headline: Student rankings higher after scoring error fixed. *The Capital Times*. Retrieved February 15, 2000, from Newslibrary database.

Myers, K. C. (2000, November 30). MCAS test scores missing for 48 Chatham students. *Cape Cod Times*. Retrieved December 1, 2000, from: http://www.capecodonline.com

Myers, K. C. (2001, May 22). State use of MCAS exams 'wrong'; The head of company running the tests said he does not think standardized tests should be used as a requirement for high school graduation. *Cape Cod Times.* Retrieved May 22, 2001, from: http://www.capecodtimes.com

National Assessment of Educational Progress – NAEP. (1998). *NAEP reading – Achievement level results for the states.* National Center of Educational Statistics. Retrieved June 16, 2001, from: http://www.nces.ed.gov/nationsreportcard/reading/statesachivis.asp

National Center for Education Statistics (NCES) (2000). *NCES to re-examine long-term writing trend data.* (draft press release), version 3.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America.* Chestnut Hill, MA: National Commission on Testing and Public Policy.

New Jersey Department of Education (2001). *New Jersey statewide testing system, May 2000, elementary school proficiency assessment: Language arts literacy, mathematics, and science.* Retrieved April 4, 2000, from: www.state,nj.us/njded/stass/index.html

New York Bar Exam: Can one question make a difference? (Winter, 1988). *FairTest: Examiner.* Retrieved February 2, 2000, from: http://www.fairtest.org

Nguyen, T. (1998, August 24). An exploration of ideas, issues and trends in education; Drop in reading scores by 9th-graders puzzles experts; Stanford 9: A decline similar to that by California students is seen in Arizona, Alabama, West Virginia and the District of Columbia. *Los Angeles Times.* Retrieved June 16, 2001, from Lexis-Nexis database.

*The No Child Left Behind Act of 2002*, Public Law 107-10, 107th Congress, 1st Session 2002.

None of the above: Test makers miss one. (1997, February 7). *St. Louis Post-Dispatch.* Retrieved April 27, 2001, from Lexis-Nexis database.

Note to readers. (2000, July 18). *Los Angeles Times*. Retrieved July 18, 2000, from: http://www.latimes.com

Nugent, K. (2002, September 28). Clinton teacher spots error in MCAS. *Worcester Telegram & Gazette*. Retrieved October 8, 2002, from Lexis-Nexis database.

NY CLS Educ. B 342, Article 7-A (2001).

O'Brien, E. (1990, June 9). Test firm flunks; 130 may try again. *The Boston Globe.* Retrieved April 27, 2001, from Lexis-Nexis database.

The Oklahoma Commission for Teacher Preparation. (2001). *Teacher candidate assessment.* Retrieved April 4, 2000, from: http://www.octp.org./candassess.html

Ore. admits to math test mistakes. (2000, April 6). *Infobeat.com.* (Oregon).

Orr, J., Butler, Y., Bousquet, M., and Hakuta, K. (2000, August 15). *What can we learn about the impact of Proposition 227 from SAT-9 scores?* Retrieved December 3, 2000, from: http://www.stanford.edu

O'Shea, M. E., and Tantraphol, R. (2001, October 23). MCAS score gains deemed suspicious. *Union-News* (Springfield, MA). Retrieved November 13, 2001, from: http://www.masslive.com

Peachey, P. (2002, May 29). Exam board makes yet another blunder by switching numbers of MPs on politics paper. *The Independent*. Retrieved June 19, 2002, from: http://Education.independent.co.uk

Pearce, K. (2000a, March 18). Mistake made in AIMS scoring. *The Arizona Republic.* Retrieved March 22, 2000, from: http://www.arizonarepublic.com

Pearce, K. (2000b, April 22). Court unmasks AIMS. State ordered to disclose questions. *The Arizona Republic*. Retrieved April 22, 2000, from: http://www.arizonarepublic.com

Pearce, K. (2000c, May 10). AIMS test has errors, teachers, students say. *The Arizona Republic.* Retrieved May 11, 2000, from: http://www.azcentral.com

Pearce, K., and Flannery, P. (2000, September 30). Glitch skews AIMS test results; Mislabeling won't affect individuals. *The Arizona Republic*. Retrieved April 27, 2000, from Lexis-Nexis database.

Pearce, K., Parrish, D., Villa, J., Baker, L., Koch, K., Go, K., and Fehr-Snyder, K. (2001, January 27). Exam scores to stay low: Poor results likely despite revamping. *The Arizona Republic.* Retrieved April 27, 2001, from Lexis-Nexis database.

Perlstein, L. (2001, June 15). Essay test's failing may hold back Md. students. *The Washington Post.* Retrieved June 16, 2001, from: http://www.washingtonpost.com

Plumberg, D. (2000, July 9). 30% of teachers failing skills test. *The Oklahoman*. Retrieved October 12, 2000, from: http://www.oklahoman.com

Policy and Evaluation Division (2000, September). *Reporting the Academic Performance Index growth and awards for 1999-2000; Media assistance packet. California Department of Education.* Retrieved June 14, 2001, from: http://www.cde.ca.gov/psaa/api/fallapi/gwmedia.pdf

Post-mortem makes sorry reading. (2000, October 13). *The Times Educational Supplement,* p. 5.

Pyramids, serpents; The exposed side. (1981, March 18). *The New York Times.* Retrieved April 27, 2001, from Lexis-Nexis database.

Qualifications and Curriculum Authority (QCA). (2001). *Standards maintained in GCSE mathematics.* Retrieved November 5, 2002, from: http://www.qca.org..uk/np/mar/standards_maintained.asp

Richardson v. Lamar Count Bd. of Educ., 729 F. sup. 806 (M. D. Ala. 1989).

Ring, W. (2000, January 19). Test grading company to pay for its mistakes. *The Burlington Free Press*, p. 3.

Ritter, K. (2002, July 30). Nevada might fine company for botched test results. *Las Vegas Sun.* Retrieved September 4, 2002, from: http://www.lasvegassun.com

Ross, J. (1997, March 5). Test flaws blamed for results. *Fort Wayne – The Journal Gazette*. Retrieved February 15, 2000, from Newslibrary database.

Sahagun, L. (1999, July 23). Report card: How L.A. county schools fared on statewide exams; What the scores tell us; Pockets of progress, but question marks, too; Bilingual Education: Experts say small increases make it difficult to tell if Prop. 227 or other reforms are responsible. *Los Angeles Times*. Retrieved November 5, 1999, from: http://www.latimes.com./archives/

Salzer, J., and MacDonald, M. (2001, June 15). Test foul-up makes schools tardy; Results being sent out a month late. *The Atlanta Journal and Constitution*. Retrieved September 4, 2002, from Lexis-Nexis database.

Samuels, C. A. (2002, August 14). Company raises Va. Test scores; Thousands were wrongly told they had failed SOL exam. *The Washington Post*. Retrieved September 8, 2002, from Lexis-Nexis database.

Sandham, J. (1998, April 1). College board rescores 15,500 tests. *Education Week*. Retrieved September 23, 1999, from: http://www.edweek.org

Schoellkopf, A. (2000, October 21). State errs in school bonuses. *Albuquerque Journal*. Retrieved October 27, 2000, from: http://www.abqjournal.com

Schrag, P. (2000, October 4). Raging reading hormones? What does the test tell? *Sacramento Bee*. Retrieved October 11, 2000, from: http://www.sacbee.com

Scoring errors sends cash to wrong schools. (2001, September 29). *Contra Costa Times*. Retrieved October 2, 2001, from Lexis-Nexis database.

Scoring settlement nets $7 million for students. (2002, November 26). *CNN.com*. Retrieved November 26, 2002, from: http://www.cnn.com/EDUCATION

Scottish Qualifications Authority, (1997). *UK Qualifications for Entry to Higher Education*. Retrieved August 31, 2000, from: http://www.ucas.ac.uk/higher/candq/ukquals/97/9/scot/right.html

Seymour, L. (2001, November 28). Va. lowers passing score for some SOL tests. *The Washington Post*. Retrieved December 28, 2002, from Lexis-Nexis database.

Shafer, L. (2000a, April 8). State will disclose school rating data. But the parent who filed the public records suit must do his own math to get 'peer' school list he sought from Education officials. *Contra Costa Times*. Retrieved April 10, 2000, from: http://www.nytimes.com

Shafer, L. (2000b, April 27). Schools get new rankings. The state's reformulated 'similar schools' ratings leave many districts disappointed with results. *Contra Costa Times*. Retrieved April 30, 2000, from: http://www.hotcoco.com

Sherwood, R., and Pearce, K. (2000, January 19). Proposal to shield AIMS test advances. Exam's integrity considered at stake. *The Arizona Republic*. Retrieved January 29, 2000, from: http://www.azcentral.com

Sheldon, D. (2002, October 21). Students in exams chaos sue en masse. *The Express* (UK). Retrieved November 3, 2002, from Lexis-Nexis database.

Siemaszko, C. (1997, February 7). Testers: SAT's the way it is, we erred. *Daily News* (New York). Retrieved April 27, 2001, from Lexis-Nexis database.

Silberman, T. (2001, May 22). High passing rate on math tests a mistake, officials say. *Newsobserver.com*. Retrieved May 31, 2001, from: http://www.newsobserver.com

Singer, D. (1998, October 3). Analysis of standardized math tests in Missouri is flawed, firm admits. *St Louis Post-Dispatch*. Retrieved February 23, 2000, from Newslibrary database.

Smith, D. (1998, July 19). Culprit for drop in test scores sought; Study: Educators are struggling to explain high school students' weak reading performance. *Los Angeles Times*. Retrieved June 16, 2001, from Lexis-Nexis database.

Smith, M. (1999, November 5). State slams firm for ISTEP+ errors. *Indiana Messenger-Inquirer.com*.. Retrieved November 5, 1999, from: http://www.messenger-inquirer.com

Smithers, R. (2001, August 31). Inquiry into exam grade inflation claims. *Guardian Unlimited*. Retrieved September 4, 2001, from: http://www.educationnews.org

Snyder, S. (1998, November 17). Test firm errs again on results; the problem has forced the school district to delay sending students' scores home. *Philadelphia Inquirer*. Retrieved February 9, 2000, from Newslibrary database.

Spielvogel, J. (2000, February 2). Educators fault data used in state school ratings; Some local districts ask recalculation of status among similar schools. *The San Diego Union-Tribune*. Retrieved April 27, 2001, from Lexis-Nexis database.

Staff. (1999, November 3). *The Andalusia Star News*. Retrieved February 29, 2000, from: http://www.andalusiastarnews.com

State blames official for school test error. (1993, March 4). *The New York Times*. Retrieved April 27, 2001, from Lexis-Nexis database.

State despairs of getting accurate test scores. (2002, August 17). *The Associated Press State and Local Wire*. Retrieved September 4, 2002, from Lexis-Nexis database.

State fines company for test errors. (1994, December 24). *The Hartford Courant*. Retrieved April 27, 2001, from Lexis-Nexis database.

State misidentified 203 schools as low-performing. (2002, July 30). *The Associated Press State & Local Wire*. Retrieved September 8, 2002, from Lexis-Nexis database.

State's schools eager for correct test scores. (1997, June 28). *The Kentucky Post*. Retrieved February 23, 2000, from Newslibrary database.

Steinberg, J. (2000, August 20). Increase in test scores counters dire forecasts for bilingual ban. *The New York Times*. p. 1.

Steinburg, J., and Henriques, D. (2001, May 21). When a test fails the schools, careers and reputations suffer. *The New York Times*. Retrieved Mary 21, 2001, from: http;//www.nytimes.com

Stephens, G. (1999, June 30). New information about 1999 Stanford 9 and STAR augmentation reporting for limited English proficient (LEP) students. *California Department of Education*. Retrieved February 16, 2000, from: http://www.cde.ca.gov

Stepp, H. (2001a, June 16). Schools boss hears diploma pleas; Seniors failed retake of test. *The Miami Herald*. Retrieved July 2, 2001, from: http://www.miami.com./herald

Stepp, H. (2001b, June 22). Crist says seniors can graduate. *The Miami Herald*. Retrieved June 25, 2001, from: http://www.miami.com./herald

Stetcher, B., Klein, S., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R., and Haertel, E. (2000). The effects of content, format, and inquiry level on science performance assessment scores. *Applied Measurement in Education, 13*;2, 139-160.

Sticht, T. (1988). *Military Testing and Public Policy: The Enlisted Corps*. Applied Behavioral and Cognitive Sciences, Inc.

Stokes, W. T., and K. E. (2000). Inside the 1999 MCAS: A close reading of the fourth grade language arts test for Massachusetts. *Currents in Literacy*. Retrieved June 16, 2001, from: http://www.lesley.edu

Strenio, A. F. (181). *The testing trap: How it can make or break your career and your children's futures*. New York: Rawson, Wade Publishers, Inc.

Students given wrong test results. (2001, March 22). *BBC News*. Retrieved March 23, 2001, from: http://news.bbc.co.uk

Study blames official for school test error. (1993, March 4). *The New York Times*. Retrieved April 27, 2001, from Lexis-Nexis database.

Sutoski, M. (1999, November 16). Statewide testing hits snags; Delays, inaccurate results found. *The Burlington Free Press*. p. 1.

Tabor, M. (1997, February 7). Student proves that SAT can be: (D) Wrong. *The New York Times*. Retrieved April 27, 2001, from Lexis-Nexis database.

Teachers' exams fears 'ignored.' (2000, August 28). *BBC News*. Retrieved September 2, 2000, from: http://news.bbc.co.uk

Test scores for some students expected by year's end or January. (2001, November 10). *The Associated Press State & Local Wire*. Retrieved June 26, 2002, from Lexis-Nexis database.

Thompson, M. (1999, September 29). Wisconsin students' test scores to be adjusted after company made errors; Most results are expected to improve. *Duluth News-Tribune*. Retrieved February 23, 2000, from Newslibrary database.

Timeline: A-level grading row. (2002, October 31). *BBC News*. Retrieved November 3, 2002, from: http://news.bbc.co.uk.

Timeline: Edexcel woes. (2002, May 29). *BBC News Online*. Retrieved June 19, 2002, from: http://news.bbc.co.uk

Tomlinson, M. (2002, September 27). Tomlinson A-levels inquiry. *BBC News*. Retrieved November 3, 2002, from: http://news.bbc.co.uk

Tully Tapia, S. (2000, September 13). Errors delay release of school test data. *The Orange County Register*. Retrieved September 13, 2000, from: http//www.ocregister.com

Two million pounds in prizes went to wrong schools. (2001, March 30). *BBC News*. Retrieved May 20, 2002, from: http://news.bbc.co.uk

University body 'sorry' for mistake. (2002, August 13). *BBC News*. Retrieved on August 18, 2002, from: http://news.bbc.co.uk

Vaishnav, A. (2000, November 29). Some MCAS scores missing: Students affected in 12 districts. *The Boston Globe*. Retrieved November 29, 2000, from: http://www.boston.com./dailyglobe

Vaishnav, A. (2001, May 17). MCAS test-taker sets history straight. *The Boston Globe*. Retrieved May 20, 2001, from: http://www.boston.com

Vaishnav, A. (2002, September 27). Teacher finds MCAS goof; 666 more students pass. *The Boston Globe*. Retrieved October 1, 2002, from: http://www.boston.com

Vance, D. A. (1997, June 26). Schools given wrong test scores. *The Kentucky Post*. Retrieved October 17, 1999, from Newslibrary database.

Viadero, D. (1999, October 20). CTB knew of problems earlier, Indiana districts say. *Education Week*. Retrieved October 19, 1999, from: http://www.edweek.org

Viadero, D., and Blair, J. (1999, September 29). Error affects test results in six states. *Education Week*. Retrieved September 28, 1999, from: http://www.edweek.org

Vogel, E. (2002, August 28). Test company agrees to reimburse parents for remedial help. *Las Vegas Review-Journal*. Retrieved September 4, 2002, from: http://www.lvrj.com

Walsh, E. (1981, March 24). Testing service is caught with its integers down. *The Washington Post*. Retrieved April 27, 2001, from Lexis-Nexis database.

Walters, J. (2001a, November 30). Computer skews school test results. *The Hamilton Spectator*. Retrieved November 4, 2002, from Lexis-Nexis database.

Walters, J. (2001b, December 6). School testing errors rampant. *Toronto Star/The Hamilton Spectator*. Retrieved May 11, 2002, from Lexis-Nexis database.

Warchol, A., and Bowers, M. (2000, November 4). Calculation error lowers some schools' SOL ratings. *Pilot Online*. Retrieved November 10, 2000, from: http://www.pilotonline.com

Waterfield, B. (2001, August 24). GSCE critics accused of 'scaremongering'. *Epolotix.com*. Retrieved November 5, 2002, from: http://www.epolitix.com

Weiss, K. R. (1998, March 25). Grading errors on SAT tests may cost some college hopefuls up to 100 points. *Los Angeles Times*. Sec. B, p. 2.

Welsh, J. (2000, July 29). Daughter's score simply didn't add up. *St. Paul Pioneer Press*. Retrieved July 29, 2000, from: http://www.pioneerplanet.com

Welsh, J. (2002, May 21). Judge limits damages in testing error. *St. Paul Pioneer Press*. Retrieved May 22, 2002, from: http://www.twincities.com/pioneerpress

Wieffering, E. (2000, August 1). Pearson buys NCS in $2.5 billion deal; The Eden Prairie-based testing company, in the midst of dealing with scoring errors, is bought by a British firm eager for access to U.S. schools. *Star Tribune*. Retrieved April 27, 2001, from Lexis-Nexis database

Wills, G. (2000). *Papal sins: Structure of deceit*. New York: Doubleday.

Wilogren, J. (1994, March 10). Errors found in test scoring of 2 Orange County schools. *Los Angeles Times*. Retrieved April 27, 2001, from Lexis-Nexis database.

Wilogren, J., and O'Reilly R. (1994, April 10). Scoring of school tests found to be inaccurate; Education: Officials concede using too small a sample for many grades. O.C. had six cases with under 25% tallied. *Los Angeles Times*. Retrieved April 27, 2001, from Lexis-Nexis database.

Womersley, T. (2000, November 4). Exam chief quits after damning report. *Electronic Telegraph*. Retrieved November 7, 2000, from: http://www.telegraph.co.uk

Woo, E. (1997, February 6). SAT error means math scores will rise for 45,000 students. *Los Angeles Times*. Retrieved April 27, 2001, from Lexis-Nexis database.

Woodward, W. (2001, June 6). Apology for A-level exam that asked the impossible. *The Guardian*. Retrieved June 18, 2001, from: http://Education.guardian.co.uk

Yettick, H. (2002, September 12). CSAP results fail the accuracy test. Oops – turns out fourth-graders did better, not worse. *Rocky Mountain News*. Retrieved October 1, 2002, from: http://www.rockymountainnews.com

Zoll, R. (1999a, September 28). State's education testing company linked to errors in test analysis/mistake may have held 8,600 students back in Tennessee. *St. Paul Pioneer Press*. Retrieved February 23, 2000, from Newslibrary database.

Zoll, R. (1999b, September 29). Tennessee says testing company had early warning on botched results. *Sacramento Bee*. Retrieved September 29, 1999, from: http://www.sacbee.com

**The National Board on Educational Testing and Public Policy**

NBETPP

## About the National Board on Educational Testing and Public Policy

Created as an independent monitoring system for assessment in America, the National Board on Educational Testing and Public Policy is located in the Carolyn A. and Peter S. Lynch School of Education at Boston College. The National Board provides research-based test information for policy decision making, with special attention to groups historically underserved by the educational systems of our country. Specifically, the National Board

- Monitors testing programs, policies, and products
- Evaluates the benefits and costs of testing programs in operation
- Assesses the extent to which professional standards for test development and use are met in practice

This National Board publication is supported by grants from The Ford Foundation and The Atlantic Philanthropies Foundation.

The National Board on Educational Testing and Public Policy
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Telephone: (617)552-4521
Fax: (617)552-8419
Email: nbetpp@bc.edu

BOSTON COLLEGE

Visit our website at www.bc.edu/nbetpp for more articles, the latest educational news, and information on NBETPP.