WHITE PAPER

Minimizing Testing Irregularities in Large Scale Assessment Programs

by Pasquale J. DeVito, Ph.D. and Michael Nering, Ph.D.

measured progress

February 2012

02.22.12-4:28

Minimizing Testing Irregularities in Large Scale Assessment Programs

This white paper was originally prepared in response to a U.S. Department of Education request for information to help states and local districts better understand existing best practices for preventing, detecting, and investigating testing irregularities.

At Measured Progress, we believe that to focus squarely on student learning requires that assessment programs be well managed, tightly controlled, and free of irregularities. However, with the ever-increasing scrutiny of and concern about assessment results, instances of cheating seem to be increasing.

Cheating scandals are headline grabbers. In fact, a Google search for "cheating on tests" yielded several high-profile stories that brought into question the integrity of various assessment programs. These references, of course, are not limited to large-scale assessments. http://www.wikihow.com/Cheat-On-a-Test identified 118 rather ingenious ways for students to cheat on tests. In fact, entire companies have been founded to help states and local districts identify instances of cheating.

> The most likely cheating perpetrators are those most vulnerable to consequences relating to poor student performance on assessments.

Generally, we believe that for any assessment program where cheating may occur, first and foremost there should be a policy document in place before any assessments are administered. This document should outline why it is important to monitor for testing irregularities and, most importantly, establish protocols for handling such cases when they arise. The policy should be developed by the government agency responsible for the program—not by the assessment vendor. It should also be vetted by various stakeholders within the assessment agency and perhaps by external groups (e.g., technical advisory committees). The policy document should articulate how these three lines of evidence should be used collectively to address cheating concerns:

- **Physical evidence:** typically during a cheating investigation some form of physical evidence is collected. This can be anything from scraps of paper to erasure marks on answer documents or even video footage.
- **Eyewitness or other direct testimony:** either someone confesses to a cheating incident or there is an accusation about cheating occurring.
- **Statistical evidence:** cheating on a test often results in an unexplainable test score or set of scores. That is, either compared to previous assessment results, or compared to a cohort of students, the student, classroom, or school results are a statistical outlier.

Cheating on a test and detecting the cheating behavior can be accomplished in a number of ways. Analysis of erasures on standardized tests has received increased attention lately, fueled primarily by instances of extensive cheating by teachers and administrators on large-scale assessments in major school districts, such as Atlanta and the District of Columbia. As the emphasis on and stakes for teacher and school accountability increase, it is reasonable to assume that instances of cheating will also grow.

It is important to note that within the context of No Child Left Behind, cheating can occur at several levels. For example, an individual student might

cheat to inflate his or her test score. However, it is more likely that cheating will occur at the classroom, school, or district level due to how the test results are used (e.g., for accountability purposes). The most likely cheating perpetrators are those most vulnerable to consequences relating to poor student performance on assessments. Any policies that are developed in an attempt to detect cheating should take this into account.

A September 13, 2011, *USA Today* survey indicated that 20 states and the District of Columbia performed erasure analysis on paper-and-pencil tests during the 2010-2011 school year. Erasure detection analysis often is conducted after the fact to "validate" that probable cheating occurred by students, teachers, or administrators erasing incorrect answers and replacing them with correct ones. As erasure detection analysis attracts more attention and popularity, requests for it likely will increase and become routine, rather than taking place only after an incident has been alleged.

Some believe that computerized testing will largely eliminate the "cheating by erasures" issues. Computer-based testing (CBT), particularly computer adaptive methodology, can substantially enhance test security; however, CBT does not eliminate erasure detection analysis as a viable tool. The detection of "erasures" simply becomes electronic. Through applications like key stroke analysis, the amount, direction, and pattern of changes to answers can be analyzed in ways similar to paper-based testing scenarios.

Erasure detection, either physical or electronic, can be a very useful tool in the overall toolkit, but it should not be seen as the primary means of detecting or preventing cheating. In fact, erasure analysis is a rather blunt tool for detecting inappropriate behavior and should be considered part of an overall best practice test security and cheating prevention strategy. Strategies for detecting or preventing irregularities on large-scale tests should include the following at a minimum:

- Strong and definitive language in administration manuals and other documents on test security procedures, responsibilities of all staff associated with the administration of the assessments, and penalties that may result from irregularities. This language reinforces the importance of maintaining security and integrity throughout the assessment process and helps to foster a district and school culture that focuses on providing accurate and informative student results.
- Processes that require relevant district and school personnel to certify that proper procedures are taken with respect to the security and chain of custody provisions necessary to ensure that the tests are not compromised.
- Random monitoring of test sites, so that districts and school personnel are aware that observations of test sessions and subsequent compilation of tests for delivery to the test contractor may occur.
- Detailed examination of school and district results over time to identify unusual patterns of gains (usually defined as three standard deviation units or greater) for the overall student population and selected subgroups. If aggregate results indicate dramatic gains from one year to the next or across a series of years, further examination should be undertaken. In most educational settings, gains in student achievement are positively incremental, so large leaps in results over time could be either an indication of a particularly effective education strategy or of artificial inflation of the scores through cheating of some sort.
- Aggressive review and investigation of any allegations of cheating or alleged collusion from students, parents, teachers, or administrators. Any and all allegations should be swiftly and actively pursued by relevant state personnel. The investigations should be the result of the guidance and instructions previously disseminated to the district and school staff administering the tests.
- Routine erasure detection analysis and specific follow-up scrutiny of suspect schools or districts. The interpretation of the data should take into account the variations that may occur at schools with very small populations.

One Erasure Detection Strategy from Measured Progress

Not long ago, a client state asked us to conduct an erasure analysis. At the time, the state was investigating serious allegations of cheating at a school and was turning to erasure detection to provide information for an upcoming court case against the school. Measured Progress designed a study that compared erasure rates from the school of interest against the overall rates in that state.

We found that there are currently no generally acceptable industry standards for conducting erasure detection analysis. After much internal discussion, we settled on an operational definition. For purposes of the study, an erasure is said to occur when there are at least two answers on the student answer sheet bubbled in for an item when: 1) at least one of the bubbles has a minimum of 10 pixels and 2) another bubble is at least 50 pixels greater. The analysis focused on providing the state agency with a dispassionate look at the data to:

- Compare the erasure rate in the school of interest to that of the whole state, and
- Compare the erasure incorrect-to-correct rate in the school of interest to the corresponding rate for the whole state.

More recently, we have conceived of ways to expand and strengthen the analysis and have implemented additional steps. Once overall and directional erasure rates for all schools have been analyzed, the state staff will use multiple methods to identify schools that may require further examination. Using computerized modules built for the Measured Progress Keyed from Image (KFI) system, we will examine enhanced electronic images of the erasures for selected schools to distinguish between actual erasures and aberrations such as stray marks, paper degradation, etc. Once this step is completed, if state agency staff decides to pursue possible legal action against a school or district, we will pull the original student answer booklets for closer examination.

> When it comes to erasure issues, testing companies should serve as objective and dispassionate data analysts, not as data interpreters or judges of the motivation of students or others.

The "Hanging Chad"

You may recall during the 2000 U.S. presidential election the incident in Florida around the "hanging chads." Because of the unusually close race between George W. Bush and Al Gore, election officials were required to do several recounts of the ballots. In fact, the race was so close that in some cases officials tried to validate the intent of the voter through a subjective visual inspection. Did they really mean to vote the way they appeared to vote? This question plagued voting officials throughout the recount process.

Detecting erasures is somewhat similar to the hanging chad problem in that you are trying to surmise something about a behavior based on little physical evidence. In using the pixel difference method discussed above, we are actually unable to make a claim that we have, in fact, detected an erasure. Furthermore, doing a visual inspection of the KFIs does not necessarily give conclusive evidence that what you are looking at is the result of an erasure. The business of detecting cheating through an erasure analysis is tricky. This is precisely why we recommend that other supportive information evidence (eyewitness testimony and statistical data) are used along with this physical evidence.

Considerations in Erasure Analyses

There are numerous factors to consider with erasure analysis:

- It is important to identify where erasure analysis fits into the overall policy plan, as well as whether the analysis takes place before or after cheating is suspected, or at both times.
- It is not reasonably possible to infer a student's intent from examining rates of erasures. It is important to consider erasure analysis to be a

dispassionate examination of the data. It is up to the assessment agency to assign meaning to the results based on other relevant information collected at the state and local level.

- Cheating scandals are widely covered in the media and often end up in court proceedings. When it comes to erasure issues, testing companies should serve as objective and dispassionate data analysts, not as data interpreters or judges of the motivation of students or others. This is a defensible and reasonable role for testing companies in cases where they are ordered to testify in legal proceedings.
- The pixel and discrepancy criteria discussed above were selected as the operational definition for determining what an erasure is. Since there is no industry standard in this area, it is reasonable to assume that not all assessment agencies embrace the same criteria. We encourage discussions around the selection of pixel and discrepancy criteria for detecting erasures.
- A key consideration is whether the analysis targets the school level, the classroom level, or both. To date, much attention has been focused on school-level results but it is reasonable to expect that cheating demonstrated through inappropriate erasures is likely to happen at the classroom level, as well. To be able to analyze the data at a classroom level, it is important to ensure that the student identifier information collected at the time of testing allows for using the classroom as a unit of analysis.
- Sample size is an important consideration. Small classrooms or grades within schools, as well as outplacement schools, are often eliminated from erasure analysis because of skewed or unreliable results. The rules for inclusion of schools should be clearly outlined prior to conducting the analyses.

As stated earlier, we at Measured Progress believe there are three main lines of evidence that should be considered to have a comprehensive and effective strategy for minimizing testing irregularities in large-scale assessment programs. These are: 1) physical evidence, 2) eyewitness or other direct testimony, and 3) statistical evidence. While it may be impossible to completely eliminate cheating on large-scale tests, sufficient attention paid to these three lines of evidence by state and local agencies and their assessment contractors can go a long way toward decreasing the number of cheating instances and ensuring the public that accurate results are being collected, analyzed, and reported.

Pasquale DeVito is a client services director at Measured Progress. A nationally recognized expert in assessment and evaluation, he served for 15 years as assessment director for the State of Rhode Island. He earned a doctoral degree in educational research, measurement, and evaluation from Boston College.

Michael Nering is assistant vice president for research and analysis at Measured Progress. He brings to his role a great depth of psychometric expertise. He plays an active role in the measurement research community. He earned a doctoral degree in psychology/psychometric methods from the University of Minnesota.