TO: P-12 Education Committee

FROM: John B. King, Jr.
Valerie Grey

SUBJECT: Testing Improvement Options

DATE: September 6, 2011

AUTHORIZATION(S):

SUMMARY

Issue for Action

How do we best ensure the integrity of 3-8 Assessments and Regents Exams in order to: (1) accurately measure student performance; and (2) ensure the continued development of our testing program into a sophisticated and rigorous next generation system necessary for meaningful education reform?

Reason(s) for Consideration

Review of policy

Proposed Handling

The question will come before the P-12 Education Committee at its September 2011 meeting where it will be discussed and action will be taken.

Background Information

Cheating scandals in Atlanta, Philadelphia and other cities around the country have fostered growing concern about the integrity of standardized test administration and scoring. Here in New York, as standardized test scores are increasingly utilized for school and district accountability and as a component of teacher and principal performance evaluations, it is imperative that those tests are not compromised. A reliable measure of student performance is vital to students' college and career preparedness. It is important to remember that the vast majority of educators approach the assessment process with integrity and are committed to providing a valid test process.

On August 1, the Commissioner announced an internal workgroup to review and recommend to the Commissioner and Board of Regents actions to reinforce the integrity of New York's testing system and create a model testing program based on best practices. The workgroup has reviewed the New York State administration and scoring processes, as well as those of other states' and cities', and has identified a number of options to augment our current processes. The actions identified include administrative, regulatory, and statutory/budget options and supplement actions already taken in recent months to improve the testing system. Improvements to the testing system should be viewed as a continuum – first addressing the immediate changes that can be made to the current system; second, revamping and reforming the system as part of the educational reform process; and lastly, planning for what the system will look like in the future.

Overview of Assessments and Exams

SED administers approximately six million exams per year in a number of testing programs: Regents/RCTs; Grades 3-8 ELA & Mathematics; Science 4 & 8; the New York State English as a Second Language Achievement Test (NYSESLAT); and the New York State Alternate Assessments (NYSAA).

SED's 3-8 assessments, NYSAA and Regents exams in Integrated Algebra, English, Living Environment and Earth Science have earned full approval in the Title I Peer Review by the USED. It is critical that SED ensure that the tests are administered and scored according to standardized procedures as outlined in the *Standards for Educational and Psychological Testing* (generally referred to as the "*Joint Standards*").

SED's Office of State Assessment oversees the development and administration of the state's assessment program including Regents exams. Local school officials are responsible for scoring the exams and reporting the results to SED. Based on knowledge of practice in several other states, conversations with testing vendors and national testing experts, New York is the only state we know of that employs local scanning and scoring. Below is a summary of how the different NYS exams are currently scored:

NYSESLAT	Exams are hand scored at the local level (open response) and scanned at the RICs and Big 5 Centers (multiple choice).
Grades 3-8 Math and ELA	Exams are hand scored at the local level (open response) and scanned at the RICs and Big 5 Centers (multiple choice).
NYS Alternative Assessment	All exams are scored regionally through the BOCES and Big 5 scoring centers.
RCT	RCTs are scored locally by teachers and not scanned.
Regents Exams	All scored locally (both open response and multiple choice). Scanning began this year and will be fully phased in next year. However, due to the tight timeframes for graduation the scanning is done after the tests are hand scored.

Current Efforts

Since 2010, SED has taken a number of steps to increase oversight of local school districts to ensure that Regents exams are accurately scored. These measures include:

- The Regents exams program has been audited by the Office of the State Comptroller twice. Most recently, an audit of Oversight of Scoring Practices on Regents Examinations was issued in November of 2009 and a follow up to that audit was just released in August 2011. The 2009 audit reviewed selected scored exams to evaluate school districts' compliance with guidelines and identified inaccuracies that tended to inflate exam scores across the State. The report included recommendations that certain actions be taken by the Department to strengthen its oversight of local scoring practices for Regents exams. The Department implemented virtually all of the recommendations. Changes included a new certification of training for proctors and the phase-in of scanning of the exams.

- Further, beginning this past year, schools were no longer permitted to rescore any open-ended questions on Regents exams following initial scoring, a longstanding practice that had caused a statistically improbable grouping of scores around the key passing marks of 55 and 65.

- Beginning in the 2011-2012 school year, a provision of the new teacher and principal evaluation regulation will require school districts and BOCES to "ensure that teachers or principals do not have a vested interest in the outcome of the assessments they score." Districts will be required to use external scoring, regional scoring, or distributed scoring technology for assessments used for evaluation purposes. This presents a unique challenge for Regents exam scoring given the quick turnaround required to score exams prior to graduation. In some parts of the state, the organizational capacity of external or regional paper-and-pencil scoring centers may exceed the amount of time available to accurately score Regents exams.

Testing Improvement Options for Administration and Scoring of Tests

SED has enhanced its processes to ensure that scoring procedures are followed along with site visits during the administration of Regents exams and additional training on proper exam scoring techniques. However, the current state assessment administrative procedures can be improved. Under the current system it has been possible for teachers to proctor assessments for students whom they teach. This situation has placed educators in an unfortunate role, where they may have been tempted to help their students as they work through assessment materials. Likewise, the current system has allowed for large testing windows, up to two weeks in some cases. Testing windows are recognized as areas for loss of item, answer, and answer document security. The system also places large administrative demands upon schools, whereby they need to maintain a completely secure site for up to two weeks.

The following administrative actions have and will be taken to prevent potential cheating and enhance the security of assessments and exams; they include:

- Requiring each grades 3-8 exam book to be administered on the same day across the state as opposed to allowing a test date window. This will create tighter controls for an answer sheet and minimize discussion about an exam. Exceptions will be made for students who are absent and students who require testing accommodations. (Note: Same day test administration for all 3-8 students may pose logistical challenges in some districts. It will almost certainly require rethinking staff deployment (e.g., role of central office staff, role of high school staff in 3-8 assessments, etc.) and approaches to accommodations (consistent with students' IEPs).)

- Requiring training certifications. Expand to 3-8 assessments the requirement that all teachers and administrators must certify that they have received and will follow security protocols for state assessments. This is currently required for Regents Exams only.

There are additional administrative and regulatory actions that could be considered to enhance the current testing system, including:

- Prohibiting teachers from proctoring exams for their own students or in their certification area. Research has indicated the prevalence of cheating increases when teachers administer exams to their own students. Restricting who can proctor could remove a temptation to help students on the exams by questioning the answers the students select or by providing tips or hints during the testing process.

- Prohibiting teachers from scoring their own students' Regents Exams and State assessments. This current regulatory requirement applies as school districts negotiate teacher and principal evaluation agreements. To ensure integrity across all State assessments, the Department could require that all school districts prohibit teachers from scoring their own student's exams. Districts would then be required to use external scoring, regional scoring, or distributed scoring technology for assessments used for evaluation purposes. A change to the Regents Exam calendar may be required to allow for districts to implement this new security provision between when the exams are administered and when graduation takes place (currently as little as a few days).

- Retaining exams for a longer period of time. Extend the records retention requirements from the current standard of at least one year to allow for improved investigations and research into potential issues.

Modernize and Improve Test Scoring and Add Cost Effective Cheating Detection Measures

Statewide Centralized Scanning and Detection

Beyond administrative reforms, several advances in scanning and scoring can also be used to ensure a more secure and valid assessment system. Under the current Regents Exam system, many schools hand score both multiple choice and open response items at the local level and only after scores have been finalized are the forms scanned by the Regional Information Centers, Big Five Scanning Centers or by the high school (who submit a file to the RIC or Big Five). For the 3-8 assessments, open responses are first scored at local or regional scoring centers, open response scores are recorded by the scorer on the same answer sheet as the student's multiple choice responses, and the answer sheets are sent to RICs or Big 5 scan centers for scanning of all items and scoring of multiple choice items. Currently, New York is the only state known to use local, hand scoring, or regional scanning of test responses.

Since 2003, centralized scanning of multiple choice item responses has been recognized as both cost effective, a best practice for ensuring test integrity and essential for rapid return of test scores. Centralized scanning procedures also allow for readily applied methods for detection of testing irregularities. Several types of statistical analyses are employed with data that are normally available in conjunction with a testing program, or with data that can be easily obtained by most programs. These analyses, in the most simple form, are often used to detect student-level impropriety and then—when combined with other methods/techniques—to detect class- and school-level irregularities. Typical types of routine statistics include:

- Erasure Analysis: Frequently used and widely accepted study of testing irregularities. Often undertaken once cheating is suspected, erasure analyses are used to check for statistically improbable rates of changed responses. If requested, most contractors can conduct erasure analyses simultaneous to the scoring process. Erasure analysis tests are used to identify statistically unlikely numbers of erasures on an individual student's answer sheet, which could be attributable to student-, class-, or school-level cheating. Such analyses are used to inform claims regarding teachers coaching students to provide a specific response during the test administration or teacher / administrator tampering of filled-in answer sheet. To conduct erasure analyses at scale, it is necessary to utilize industrial scanners that can detect and record gradations of eraser marks on individual test answer sheets.

- Aberrant Response Analyses: These analyses are used to investigate the reasonableness of an examinee's answers to a set of test items (e.g., when students of low ability respond to items of greater difficulty more successfully than their ability would suggest). Aberrant Response Analyses are often readily available and easily calculated at the time of score estimation. Possible inferences stemming from these analyses include student- and educator-level cheating.

- Very Similar Test Response Statistics: This family of analyses is used to unearth instances where two (or more) students' test responses are more similar than statistically expected. The main inferential goal of this analytic path is the detection of copying. When expanded to investigate the similarity of more than two response sets, inferences regarding student collusion or possible teacher tampering can be made. One specific type of test for similarity, Spurious String Analysis, is used to detect unlikely blocks of student responses within the overall answer form. This type of analysis was used by Leavitt and Jacob in Chicago Public Schools and outlined in the book *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything.* [1]

- Score Fluctuation: These analyses are used to discern unusually large gain in cores from year to year. These tests leverage expected change from one year to the next to identify instances in which scores rise greater than those suggested by a given statistical model. These tests are used to inform a variety of possible teacher-level irregularities (e.g., when an entire class's scores fluctuate in an unlikely manner). Similarly, score fluctuation analyses could also leverage Student Growth Percentile (SGP) analyses to identify aberrant score fluctuations.

A major consideration for most states in the move to centralized scanning has been the standardization and security measures that contractors employ in the process. Without standardization and security, it is difficult, overly time-consuming, and expensive to undertake cheating analyses. (Please refer to Appendix A for an explanation of centralized scanning).

It is recommended that the Board direct the SED workgroup to take the following actions and report back to the Committee in October:

- Develop a statewide centralized scanning proposal. Establishing a statewide system of scanning is expected to have a substantially lower overall cost than local scanning due to economies of scale. In order to develop a statewide scanning system the State must identify a funding mechanism for it. A preliminary review of other states' costs indicates that centralizing would likely lower overall costs significantly.

- Include the purchase of erasure analysis in the centralized scanning proposal. The use of a centralized system of scanning will allow for the addition of erasure analysis at nominal cost.

- Evaluate the inclusion of other statistical analyses such as aberrant response, very similar response and score fluctuation that should be included in the centralized scanning proposal. Create a group of expert advisors, with the advice of the

[1] Levitt, S.D. & Dubner, S.J. (2005). *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. New York: William Morrow.

Technical Advisory Group, to assist with the selection of data forensic services that improve our ability to identify and adjudicate potential instances of cheating.

Distributed Scoring

Distributed scoring uses online services to break down the physical barriers that have been historically associated with open responses scoring in large-scale assessments. In practice, distributed scoring includes the scanning of responses, anonymously warehousing student work, and then distributing responses digitally to scorers, enabling experienced scorers to effectively score assessments from afar. Distributed scoring allows the use of a large pool of qualified scorers, increases the integrity of scoring, and allows a move away from local scoring of open-response items. (See Appendix A for a description of the relationship between centralized scanning and distributive scoring).

Distributed scoring is currently recommended as best practice by the Congressional Budget Office due to both cost mitigating factors and the inherent advantages in deterring test tampering.[2] As a cost saving mechanism, distributed scoring leverages scale in scanning of responses and takes advantage of technology to distribute responses to qualified scorers. By making scoring anonymous, and distributing responses at random throughout the state or country, the possibility for vested interest and temptation to tamper is negligible. It is also a tool that can be utilized for professional development.

It is recommended that the Board direct the SED workgroup to take the following actions and report back to the Committee in October:

- Research the potential to integrate the use of distributed scoring. This would assist local districts and further improve the testing system.

Multi-State Consortium (PARCC) and the Future of Testing – Online (longer term)

In January, 2010, the Regents endorsed the participation of New York State in the 24-state Partnership for the Assessment of Readiness for College and Careers (PARCC). PARCC is a consortium of states that worked together on a joint proposal to USDE to seek Race to the Top funding for the development of a K-12 assessment system aligned to the Common Core State Standards in English language arts and mathematics for grades 3 - 11. PARCC was awarded a total of $185 million in September 2010.

One of the hallmarks of the PARCC design is that it is a computer-based test (with some exceptions for paper-based administration in the elementary grades). PARCC is working diligently to ensure that as much of the assessment as possible can be machine-scored. For traditional multiple-response items this task is simple and readily achieved. Likewise, for shorter, open-responses PARCC is committed to pursuing the

[2] Lomax, E.D. (2010) State Assessments Required by the NCLB Act: An Analysis of Requirements, Funding, and Costs. Congresional Research Service: Washington, DC

use of artificial-intelligence, based scoring methods. For rating full-length, essay responses it is PARCC's belief that the increased demand will spark innovations in text recognition programming.

Given the number of states currently pursuing computer-based assessment platforms, with both PARCC and SBAC committed to online administration, technological capacity looms large as a potential issue. While computer-based assessment has been recognized for its advantages, in regards to cost and measurement functions, it seems imperative for USED to aid schools in building the necessary technological capabilities.

Follow up on Incident Reporting

SED's Office of Assessment Policy, Development and Administration is the central point of contact for reporting irregularities in the administration and/or scoring of State exams. Communications range from administrator reports of student fraud, misadministrations and wrong-doing by professionals to anonymous allegations of fraud and cheating by professionals. Information is received from many sources including parents, teachers, school building and district administrators via telephone, email, fax, and through the NYSED's Fraud, Waste and Abuse Report hotline.

It is recommended that the SED engage an independent entity to review the existing intake and response process for testing irregularity information and provide recommendations for improvement. The independent entity should have proven knowledge attained through experience developing and managing incident reporting systems. A complete evaluation of the existing system and follow-up methods would be reviewed.

Recommendations

SED will undertake the administrative actions in the testing process as outlined above including:

- Requiring grades 3-8 exams to be administered on the same day. This action was taken in the testing calendar released on August 26, 2011.
- Requiring a training certification for proctoring and scoring 3-8 assessments similar to what is required for the Regents exams.

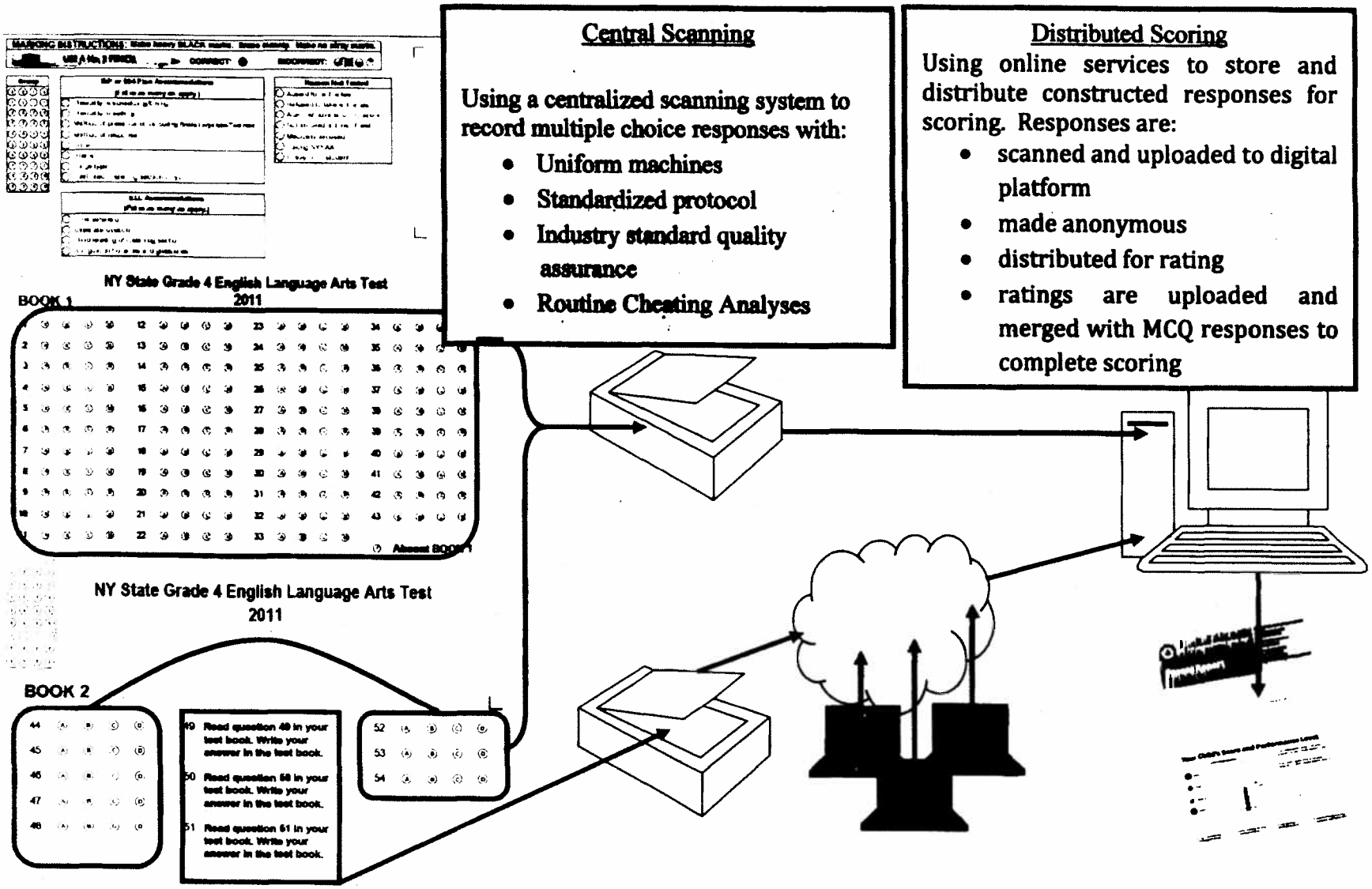
.The following action is before the Board for approval:

- The Department shall take immediate action to secure an independent review of the Department's procedures related to incident reporting and follow up of allegations of testing impropriety at schools.

The Board directs Department staff to further develop specific proposals for consideration in October in the following areas:

- The Department could require that all school districts prohibit teachers from scoring their own student's State assessments.
- Required that districts retain assessment and exam answer sheets longer than one year.
- Prohibiting teachers from proctoring exams for their own students or in their certification area.
- Centralized statewide scanning and scoring of multiple choice assessments and exams that would include utilization of erasure analysis and enhanced error pattern analysis and data forensics.
- Development of a distributed scoring platform that would be used to score open responses throughout the State.

Appendix A: Central Scanning and Distributed Scoring



Central Scanning

Using a centralized scanning system to record multiple choice responses with:

- Uniform machines
- Standardized protocol
- Industry standard quality assurance
- Routine Cheating Analyses

Distributed Scoring

Using online services to store and distribute constructed responses for scoring. Responses are:

- scanned and uploaded to digital platform
- made anonymous
- distributed for rating
- ratings are uploaded and merged with MCQ responses to complete scoring