

STATE ASSESSMENTS & DATA INTERPRETATION

QUESTIONS & ANSWERS

March 2006

This question and answer document was the result of a Professional Learning Community on state assessments and data interpretation which included representatives from Cayuga-Onondaga BOCES, DCMO BOCES, OCM BOCES, ONC BOCES, and Oswego BOCES. The questions were brought to the table from component schools and were researched and discussed by the group. Kimberly O'Malley, Lead Psychometrician from The Pearson Group assisted in the editing of this document. The document should be considered a work in progress and your suggestions and additional questions would be appreciated.

Cayuga-Onondaga BOCES

Becky Kaune- Director of Curriculum and Staff Development - bkaune@cayboces.org

Leela George- Coordinator of Data, Curriculum & Assessment - lgeorge@cayboces.org

OCM BOCES

Jessica Cohen- District Superintendent - jcohen@ocmboces.org

Pat Richards- Assistant Superintendent for Instructional Support Services -
prichard@ocmboces.org

Peter McCarthy- Staff Development Specialist - pmccarth@ocmboces.org

Oswego BOCES

Maryann Preston - Research Specialist - MPreston@oswegoboces.org

DCMO BOCES

David Blom - Director of Instructional Support Services - blomd@dcmoboces.com

Marki Clair - Assistant Superintendent for Instructional Services - ClairM@dcmoboces.com



Note: A number of updates to this version of this document have been included to reflect changes to the 3rd through 8th grade NYS Assessment system as of the 2005-06 school year. These changes are indicated by the boxed text.

What is the scope of the New York State assessment system?

The New York State Assessment **SYSTEM** is more than just a series of tests. This system provides guidance for effective school planning in the following areas:

1. **LEARNING STANDARDS:** The content and skills students need to know and be able to do.
 - a. **ELA-** Students will be able to read, write, listen and speak for information and understanding
 - b. **MATH-**Students will communicate and reason mathematically
2. **PERFORMANCE INDICATORS:** How students demonstrate their content knowledge
 - ELA-** Identify main ideas and supporting details
 - MATH-**Describe objects, relationships, solutions and rationale using appropriate vocabulary
3. **ASSESSMENT:** A measurement of level of performance
4. **ACCOUNTABILITY:** How the results of schools and districts are measured against federal and state yearly progress targets(AYP)

How does the state ensure that the tests are good? NYS has developed and adopted a 19-step examination development process. These steps ensure that the tests are valid, reliable, free of bias, and meet the highest standards for technical quality. In addition, the tests are developed with the participation of New York State teachers.

What is the role of New York State teachers in the state assessment development process?

Teachers may be invited to participate at 8 critical points in the process, including:

1. Item Writing
2. Standard-Setting
3. Defining levels of performance
4. Reviewing content
5. Designing Pre-Tests and Field Tests
6. Scoring Pre-Tests
7. Final Eyes Review
8. Grades 3-8 Test Specification and Framework

How can local teachers become involved in the process?

Administrators and teachers can get information on how to apply in two ways:

http:// www.emsc.nysed.gov/osa/

E-mail: emscassessinfo@mail.nysed.gov

What is a field test?

A field test is given to establish the technical quality of test questions. Field tests are given to a representative sample of students. Data from those students are used to gain information about item difficulty based on student performance. Field tests are usually given on multiple forms with each form having fewer items than operational tests. Administering different forms of field tests permits all items to be tested in shorter sessions without any group of testers taking long field tests. In addition, administering different forms of field tests prevents students from being exposed to a full set of items that might end up on an operational test.

SED Field Test Matrix- <http://www.emsc.nysed.gov/3-8/fieldtest.htm>

What is an operational test?

An operational test is the final version of the test given to all students during regular testing time.

How does field-testing affect the operational test?

High failure rates on the 2003 Math A exam were in part affected by non-representative participation in field-testing. It appears that the field test population included a disproportionate number of higher performing students, which affected the scaling process. This resulted in lower scores for the students taking the test.

To avoid the risks mentioned in the quote from the Math A report above, SED developed a new (2004-05) sampling design for elementary, intermediate, and Regents field-testing. The new system depends on participation of all schools and the full range of eligible students. This stronger sampling plan will help ensure that field test samples are representative of New York State's student population.

How does the issue of readability affect test questions?

Readability formulas provide some useful information about the reading difficulty of a passage or stimulus. However, a readability score is not the only indicator of grade-level-appropriateness. Prospective passages are submitted for review to panels of New York State educators familiar with the abilities of the students to be tested and with the grade-level curricula. The passages are reviewed for readability and also for appropriateness of content, potential interest level, quality of writing, and other qualitative features that cannot be measured via readability formulae.

Any standardized test will include questions with a range of difficulty, so applying a readability formula to a question or set of questions does not demonstrate the readability of the test.

How do test developers work to ensure that the test is free of bias?

As part of the test development process, item writers are directed to avoid:

- Stereotyping based on demographic characteristics: gender, socioeconomic status, ethnicity, etc.
- Use of terms or concepts that might advantage or disadvantage any segment of the student population

In addition, items are screened for bias by a statistical analysis of field test results.

How do testing accommodations affect testing validity?

The purpose of accommodations is "to level the playing field" and allow for participation of eligible students on an equal basis to their peers. * Standards are for all students; therefore an assessment system needs to measure the achievement of all students. When the appropriate testing accommodations are in place and administered correctly, the accuracy and validity of the scores should not be compromised.

*When the use of an accommodation would change the purpose of the test and yield inaccurate scores it would not be permitted.

For example: If the purpose of the test is to determine reading ability, reading the test to the student may skew the results; however, reading the directions to the student should not alter the test purpose.

How are students' scores determined? (raw scores, scale scores, and performance levels)

- Raw scores are the total number of points earned.
- Scale scores are the statistical conversion of the total number of points earned (raw score) to a range of values representing student ability.
- Students are placed into one of the four performance levels based on their overall performance on the test as determined by their scale scores.

Raw scores, which are the total number of points earned on multiple choice, constructed response and extended response questions are converted to scale scores, which are then translated into performance levels.

For example, a 32 raw score might be converted to a 697 scale score.

Conversion charts convert raw scores into scale scores.

- For some tests such as the Global History and Geography Regents, a matrix type chart is used to convert the combination of scores on parts of the tests to a single scale score. For example, if a student received a 7 on his/her essay and a 45 on part I and part IIIA, the scale score would be 83. (August 2004)
- For other tests such as ELA 4, a linear chart is used to convert a total raw score to a single scale score. A 40 raw score yields a 747 scale score. (February 2005)



Please Note: Due to the extensive data collection and psychometric requirements of this new testing program, the time frame for issuing score reports will be different from the time frame for issuing score reports for the current grade 4 and 8 tests. Score reports for the ELA tests are scheduled to be issued to schools in early August 2006. Score reports for the mathematics tests are scheduled to be issued to schools in late September 2006. These time frames are based upon the need to collect sufficient statewide data to perform standard setting and scaling prior to the release of final student scores. These necessary psychometric functions will take approximately 140 days once scores have been submitted to SED and CTB for analysis and verification. However, these time frames will decrease for the subsequent years of the Grades 3 – 8 Testing Program, as the complete psychometric analysis and standard setting need not be repeated after the first year of test administration. Additional information on the impact of these time frames on academic intervention services and other instructional decisions will be forthcoming.

What is a scale score?

A scale score represents a statistical conversion of total number of points earned (raw score) to a range of values representing student ability.

For ELA/MATH elementary and intermediate the scales vary slightly but are generally in the range of 450 – 850. The Regents examination scale ranges from 0 – 100 but this should not be confused with a percentage of correct answers. Because the scale score range is broader than the raw score range, not every scale score in the range is possible for a given administration.

For example, raw scores on ELA 8 2004 for level 4 range from 39-43 – 5 score points – which will be converted to 5 score points on the scale ranging from 737-830.

Why do we use scale scores?

Scale scores allow comparison of performance from one year to another at the elementary and middle school level, and from one administration to another at the high school level. For example, scale scores allow comparing grade three ELA scores one year to grade three ELA scores the next year.

SED has advised that comparison of scale scores across different grade levels cannot be interpreted as meaningful.

Example: A scale score of 432 on the fourth grade ELA cannot be determined to be a better score than a scale score of 415 on the eighth grade ELA .

How is the scale created?

Using information gained through field testing and operational testing about the difficulty of items, psychometricians establish the scale.

Every point on the scale score continuum should be related to the student's status with respect to achieving State Learning Standards. In fact, this is accomplished on the State examinations by deriving those scale scores so that each score is related to a point on the array of skills that comprise State Learning Standards. In this way, a student at any given scale score can be said to have an expected probability of having any of the skills or knowledge that make up State Learning Standards. For example, for a given skill, e.g., plotting points on a graph, higher scale scores indicate a higher probability that the student has mastered that skill, while lower scale scores indicate a lower chance. More importantly, this probability can be precisely computed from the student's scale score, so that educators and teachers can derive from the scale score just which skills are more likely to be within the child's capability. Other kinds of scores, such as percentages correct, do not have this built in reference to achievement, and therefore are less valid measures of achieving State Learning Standards. (Gerald E. DeMauro, SED July 2002)

How do we determine the cut scores that separate the levels of performance (e.g., 1,2,3,4)?

In setting the final performance standards, the Regents use the recommendations from the standard setting committee which includes New York State teachers and other

professionals. The standard-setting committee meets prior to the time when scores on the first administration of a test are reported. The committee typically follows an item mapping, or bookmarking, process for recommending cut scores to the Regents. The process includes describing what students should know and be able to do to reach the different performance standards, reviewing the questions on the test form, and making several rounds of judgments about where the cut scores should be placed. During the item mapping process, committee members review the test items in order of difficulty. They decide which questions, and to what extent on open-ended questions a student would have to answer to meet the performance standards. The committee's recommendations are then reported to the Regents for the final policy decision.

Can all students pass the test?

New York State assessments are not norm-referenced; they are criterion referenced. There is no State manipulation of test results to have a certain percentage of students pass or fail. Rather, the object is to operationally define what it means to achieve the learning standards. Theoretically, all students could reach proficiency on any given state assessment.

Can I expect equal difficulty from one administration of the test to another?

An 8th grader in January 2003 received a raw score on ELA of 31, which translated to a Level 3 score. An 8th grader in January 2004 received the same raw score, but this score translated to a Level 2. What does this say about the two tests and the two students?

The difficulty level of the test may change from one administration to the next. What does not change from test administration to the next is the difficulty of the performance standards. Students are held to the same standards on every test administration. Different forms of a test may not be equally difficult, because the group of questions selected for a test form may be slightly more or less difficult than those on another test form. Therefore, in order to maintain fairness to all students, a student who takes a slightly less difficult form of a test (e.g. January 2004) will be required to answer 1 or 2 more questions correctly. Conversely, to keep the standards at the same difficulty level, the raw score may go down on one test so that a student who takes a slightly more difficult form of a test (e.g. January 2003) may be required to answer 1 or 2 fewer questions correctly than on a slightly easier form.

Every effort is made to have different forms of a test be of equal difficulty. However, slight changes in difficulty of different tests lead to slight changes in the number of questions a student needs to answer correctly to reach a performance standard. These slight changes do not mean that the level of expected proficiency for students has changed. The changes are made so that all students are held to the same level of expected proficiency.

In the example above, both of the 8th grade students achieved 31 raw score points. Since the 31 points put the student in 2003 in a higher proficiency level than the 8th grader in 2004, the student in 2003 took a harder test than the student in 2004. The

student in 2003 was slightly more proficient than the student in 2004 because that student reached the higher performance level.

What is equating?

Equating is a procedure for measuring and controlling for variations in the difficulty of different administrations of the same test.

For example: Elementary ELA- is this years test harder or easier than last years? We need to equate the tests so that the difficulty level of the performance standards is the same. This allows all administrations of the test in a grade and content area (e.g., Grade 4 Reading) to use the same scale year to year.

Equating is the statistical process of adjusting test forms that have different difficulties. Whenever test forms that are supposed to cover the same content do not contain the exact same questions, equating must be done so student scores can be fairly compared. Remember that test forms may differ slightly in the difficulty of the specific group of questions they contain. To adjust for the difference in difficulty of the two tests, we might expect students to do better on the easier test or get fewer questions right on the harder test. Equating adjusts for these differences and ensures that all students are held to the same standard.

Why equate?

Equating is done to make the performance standards on two different administrations of a test in a specific grade and content area equivalent year to year.

Example: 125 questions correct on one test may convert to the passing scaled score of 500 while on the easier test 130 questions correct may convert to the same scaled score of 500. Even though the number of correct questions may convert to a different scale score on different test administrations, the performance a student needs to demonstrate to reach the different performance levels (e.g., passing) will remain the same year to year.

What is the purpose of horizontal scaling?

Horizontal scaling is a particular application of equating that relates tests in the same grade and of similar content from year to year. Most commonly this occurs when equating multiple forms of a test given at a particular grade level for example the ELA 4 test each year. Horizontal equating is also used for Regents exams to ensure that different forms of the test (e.g. January and June versions) provide comparable opportunities for students to pass. Another interesting use of horizontal equating is to help ensure that test versions translated into different languages equate to each other and to the English version.

What is the purpose of vertical scaling?

Vertical scaling attempts to use a single scale to summarize student achievement across grade levels. An example of this is the scale used by the TerraNova K-12. There are several assumptions that must be satisfied in a vertical scaling model. Most importantly, the tests must test the same general content or concept. For this reason, the scales are often developed across two adjacent grades through overlapping test

items. Vertical scaling is used mainly in reading and mathematics where content and concepts form a reasonable continuum across the grade levels. When a test has a vertical scale, it is possible to compare a student's grade 3 score to her grade 4 score. These vertical scales help describe students' growth in subjects from year to year.



Vertical scaling was originally considered for the new (2005-06) generation of NYS Grade 3-8 Assessments. It has been determined that this method was not practical given the time and size constraints of the test program. As of Fall 2005, SED is investigating another methodology called Vertically Moderated Standards to provide data about achievement across grade levels.

What is the relationship between scale score and passing score?

In 2003 John answered 52 out of 80 questions correctly (65%) yet he received a scale score of 63. Why didn't he pass the test?

In 2004 Mary answered 51 out of 80 questions correctly (64%) yet she received a scale score of 65. Why did she pass the test?

NY State assessments are not scored by the percent of questions answered correctly. Instead the raw score (total number of points earned) is converted to a scale score to equate different versions of the test given.

The raw score converts differently each year based on the statistical analysis of the difficulty of questions.

Mary and John's scaled scores reflect how they would have performed had they taken the test in the same year, i.e. Mary showed a higher level of mastery of the standards assessed on the test even though her raw score was lower.

Note that the Regents Examinations now use a scaled score on a scale of 100, with a scale score of 65 as the passing score. This allows Regents Examinations to be equated from one administration to the next (*January - June*), as well as from year to year.

Note: New mathematics Regents examinations will be implemented over a three year period. The Integrated Algebra test will first be administered in June 2008, Geometry in June 2009, and Algebra 2 and Trigonometry in 2010. It is very important to know that these tests will be post-equated. That is, scale scores will not be available until the process of equating has been completed after the test administration. This period is estimated to be about six weeks. Therefore, for those specific test administrations, districts may have to plan for other ways of making decisions about such issues as student placement for September, summer school requirements, and final course grade determinations.