

September 14, 2009

Dear Educator,

In the memo below you will find answers to some of the most frequently asked questions about the 2008-09 Acuity Predictive results for grades 3–8. Throughout the coming school year we will develop additional resources to explain the Acuity results in more detail and help you connect those results to practice in your schools. Please look for more information and professional development opportunities throughout the year on the [Periodic Assessment Website](#).

As always, if you have any questions or would like to discuss this further, please contact the Periodic Assessment Team at (212) 374-5162 or [periodicassessment@schools.nyc.gov](mailto:periodicassessment@schools.nyc.gov).

Best regards,

The Periodic Assessment Team

\*\*\*\*\*

**Why did the 2009 Spring Predictive assessment not report predicted performance levels on the State test?**

- A In order to develop prediction, we compare students' performance on the Acuity Predictive Assessments to their performance on the subsequent State tests. The Acuity Spring Predictives were first administered in Spring 2008 and needed to be compared to State tests that were administered in January and March 2009. The results of the 2009 State tests were not available early enough to complete the comparison to the 2008 Predictive assessments and inform prediction on the June 2009 assessments.

**What is a "good" scaled score for each grade level? If you are looking at a student longitudinal report, how much growth shows "significant gains"?**

- A There can be several ways of defining a “good score” for a student. In general, we might consider the Acuity scaled score that shows a student is very likely to score a Level 3 on the New York State test to be a good score. At an individual student level, an educator might want to look at scores and predictions in reference to the student's starting point or compared to others in the class or grade. If, for instance, a school uses an alternate pacing and no students in a particular grade have learned a group of skills represented on the Predictive Assessment, it would be important to take that fact into account when reviewing assessment results (including predictions and scaled scores) for students in that grade.

Similarly, “significant gains” can have several meanings. In the scientific sense, a significant gain for a student is a gain that is large enough that it is very unlikely to be

caused by chance or measurement error (and, therefore, shows true improvement). You could say that a student made *statistically significant* gains if the standard error of measurement (SEM) confidence intervals surrounding two scores obtained at different times did not overlap (see the following question for more information on SEM). So, if a student achieved a scaled score of 400 on the Fall 2008 Predictive with an SEM of 60, and that same student scored a 550 on the Spring 2009 Predictive with an SEM of 50, the fact that the two SEMs did not overlap would mean that the score gain is beyond what could be accounted for in measurement error.

On the other hand, you might define *educationally significant* gains relative to other benchmarks, such as bringing a student predicted to be in Level 2 on one assessment form to Level 3 on another assessment form in the future.

The Periodic Assessment Team is developing resources and strategies for schools to use to measure growth using scaled scores. In the meantime, you can use ARIS report 4 to look at the average growth by grade citywide in Acuity scaled scores from the Fall Predictive to the Spring Predictive and compare this to your student, class, or school. Alternatively, you can use ARIS report 1 to look at average Acuity scaled score by grade citywide on the Spring Predictive.

### **What is the standard error of measurement (SEM)? What affects the size of the SEM?**

- A The SEM is related to the reliability of an assessment form; it indicates the accuracy of the students' scores depending on where those students fall in the range of possible scores. The SEM tends to be smallest toward the central part of the score distribution. This is by design; in test measurement and research, assessments are frequently constructed to be most accurate at the part of the distribution of scores that most students are expected to achieve.

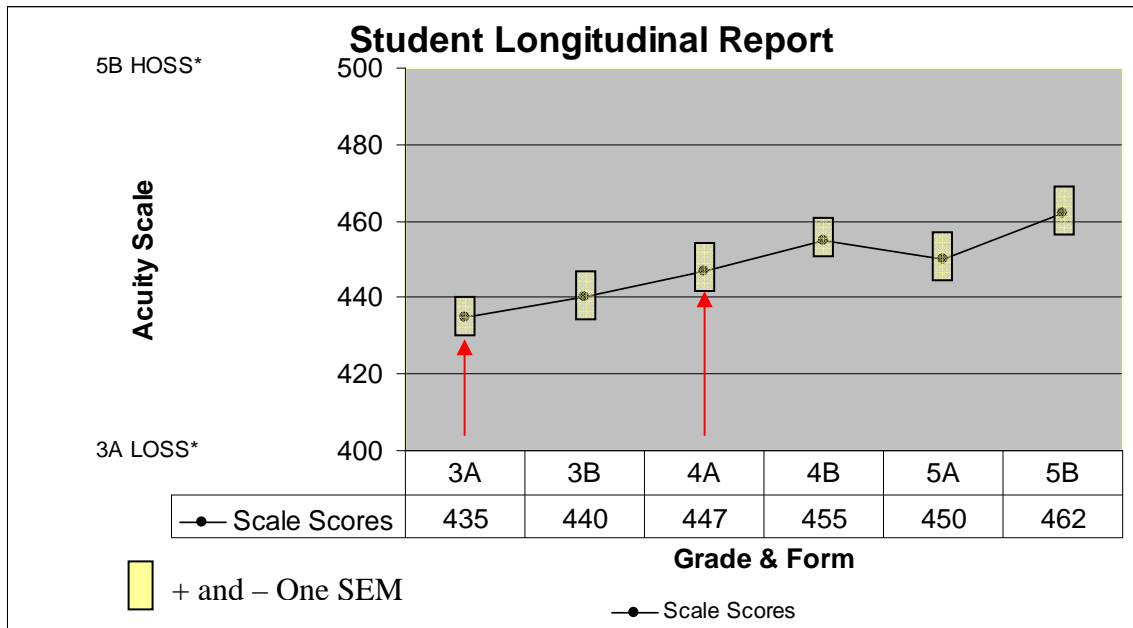
Generally, the SEM will be largest at the lowest and highest obtainable scores. This is because at the lowest end of the scale the assessment shows more about what a student does not know than what he or she does know—which does not provide a very accurate measure. What we learn about students scoring at the lowest obtainable score on any given assessment is that those students would be better measured by an easier assessment. Similarly, at the highest end of the scale students have demonstrated strong mastery of the content of that particular assessment, but may be able to demonstrate much more. In order to get more accurate scores (and lower SEMs) for very high performing students, those students would need to take a more difficult assessment.

### **What do the Acuity Predictive scores from 2008-09 tell us about growth?**

- A Students' scaled scores can be used to measure growth over the course of the school year at the student, class, school, district, or city-wide level. An increase in scaled score from the Fall Predictive Assessment to the Spring Predictive Assessment

indicates growth. Growth from one point in time (fall) to another (spring) that is larger than the SEM around each score is “statistically significant.” The standard error of measurement appears on the student longitudinal reports (available in the Reports section of Acuity) so that it can be easily observed, as illustrated below.

Note that the student’s growth represented below from Form 3A (grade 3 Fall Predictive) to Form 3B (grade 3 Spring Predictive) is positive, but the SEM bands overlap. This does not mean that growth has not occurred. The student has demonstrated 5 scaled points of growth from 435 to 440 from 3A to 3B; however, the growth was not large relative to the SEM. On the other hand, the growth from 3A to 4A (grade 4 Fall Predictive), which is also positive, is such that the SEM bands for 3A and 4A do not overlap. In this case, one can assume the increase in scores represents real student growth exclusive of measurement error.



\* HOSS - highest obtainable scaled score  
 \* LOSS - lowest obtainable scaled score

**Why is the grade-level-specific Acuity scaled score minimum and maximum range so wide? (i.e. How could a student earn the same scaled score on a 3rd grade assessment as on an 8th grade assessment?)**

- A Unlike on the New York State tests, all the Acuity Predictive Assessments for grades 3 to 8 are on a single, vertical scale in order to allow schools to compare scores and see student progress within and across years. However, although a student in grade 3 may have the same scaled score as a student in grade 8, this does not mean the two students have the same skills. For example, a student at the 90<sup>th</sup> percentile in grade 3 may have the same scaled score as a student at the 10<sup>th</sup> percentile in grade 8, but the curricula of the two grades is obviously very different—the students do not have the

same skills and it should not be claimed that the 90<sup>th</sup> percentile student in grade 3 is performing at a low grade 8 level.

In addition, a student who has a scaled score of 500 on every Acuity Predictive Assessment from the Fall of grade 3 through the Spring of grade 8 will be falling behind his or her classmates. The scale of possible scores for each grade significantly overlaps every other grade, but the lowest obtainable scaled score (LOSS) and the highest obtainable scaled score (HOSS) increase slightly from grade to grade—meaning that if a student scores the same exact scaled score each year, that scaled score will represent lower achievement relative to grade-level expectations each successive year.

The range of scores that comprise the highest probability of performing at each of the four predicted levels is also different in each grade. As a result, there is not a single point value of growth that represents “one year’s growth” from one Predictive Assessment to the next. However, a student who falls within predicted Level 3 score range in grade 4 on the Predictive Assessment and then falls within a predicted Level 3 score range in grade 5 on the Predictive Assessment has likely demonstrated a year of growth.

The complete score range across all grades is determined, following standard assessment design practice, by designating a fixed mean and standard deviation at a central grade (in this case, grade 5 or 6). The score ranges for each individual grade are then designated based on the measurement properties (i.e. difficulty of the questions, placement of questions on the test, etc.) of the individual grade-specific assessments within the context of all the assessments. The possible score ranges for each grade are wide so that the assessment can provide scores that represent the wide range of possible student performance at each grade level.

### **How predictive were the 2008-09 Fall Predictive Assessments?**

- A Citywide across grades 3–8, the Fall Predictive Assessments predicted the correct performance level on the State tests for approximately 75 percent of students. For the 25 percent of students for whom the Predictive Assessments did not accurately predict their level on State tests, the majority scored higher on the New York State tests than was predicted. There are many reasons why students’ predicted and actual performance levels could be different. Perhaps the two most likely explanations are: 1) some students did not try as hard on the Predictive assessment as on the actual test, so the Predictive assessments could not capture their actual ability or 2) after the Predictive assessments, students received targeted instruction from their teachers that allowed them to make greater than average gains in learning between the Predictives and State tests.

You can use ARIS Report 1 to look at how students across NYC were predicted to perform and Report 5 to see how accurate the predictions were citywide.

### **What action can be taken based on the Predictive Assessment data?**

- A Individual student data is useful to assess the effectiveness of instruction with individual students, for grouping of students, and to track growth over time. In addition to using the performance level predictions to determine which students are at risk of not learning all the standards assessed on State tests, educators can find valuable information about each student's strengths and identify opportunities for growth by analyzing skill level information.

Group data is best used for program evaluation at the class, school, or district level. Group data can be used to evaluate the overall level of growth in a class or school and can be used to assess the success of a particular instructional method used in the classroom or school. These data should not be used to make high-stakes decisions at the individual student, class, or school levels.

### **Which data should be used for individual students? Is some data more meaningful in aggregate?**

- A The data is more reliable at the group level because the standard error of measurement tends to cancel out at the aggregate level, leaving an educator with a better representation of average performance. However, a student's observed scaled score is the most accurate estimate of that student's true score and can be used to measure growth. The standard/strand, performance indicator and skill level information from the predictive assessments supports instructional decisions at all levels.