Test Analysis, Operational Scaling and Scoring

Test Analysis

IRT item parameter estimates were used to generate test characteristic curves (TCCs), test information functions (TIFs), and conditional standard errors of measure (CSEM). These indices were computed for each of the base forms, form-to-form linking items, and entire item pool. Figure 1 shows the overlaid TCC plots for Form A, B, linking item set and the entire item pool for grade 5. The TCC and TIF values were divided by the total number of score points for each form so that the curves can be plotted on the same scale. Figure 2 displays test information curves for Form A, B, linking item set and the entire item pool. Figure 3 illustrates the conditional standard error of measurements for the four tests.

Figure 1. Test Characteristic Curve of the Grade 5 Science Test

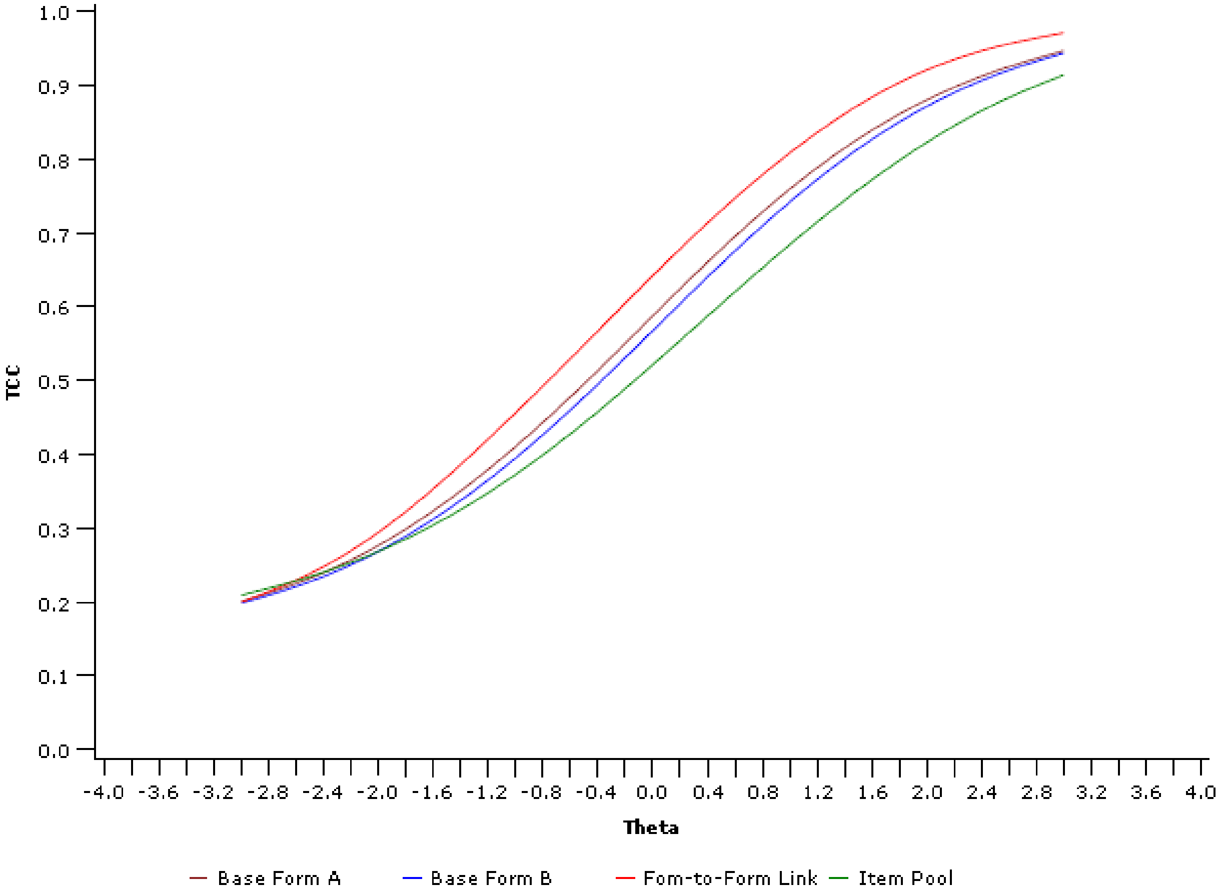


Figure 2. Test Information Function of the Grade 5 Science Test

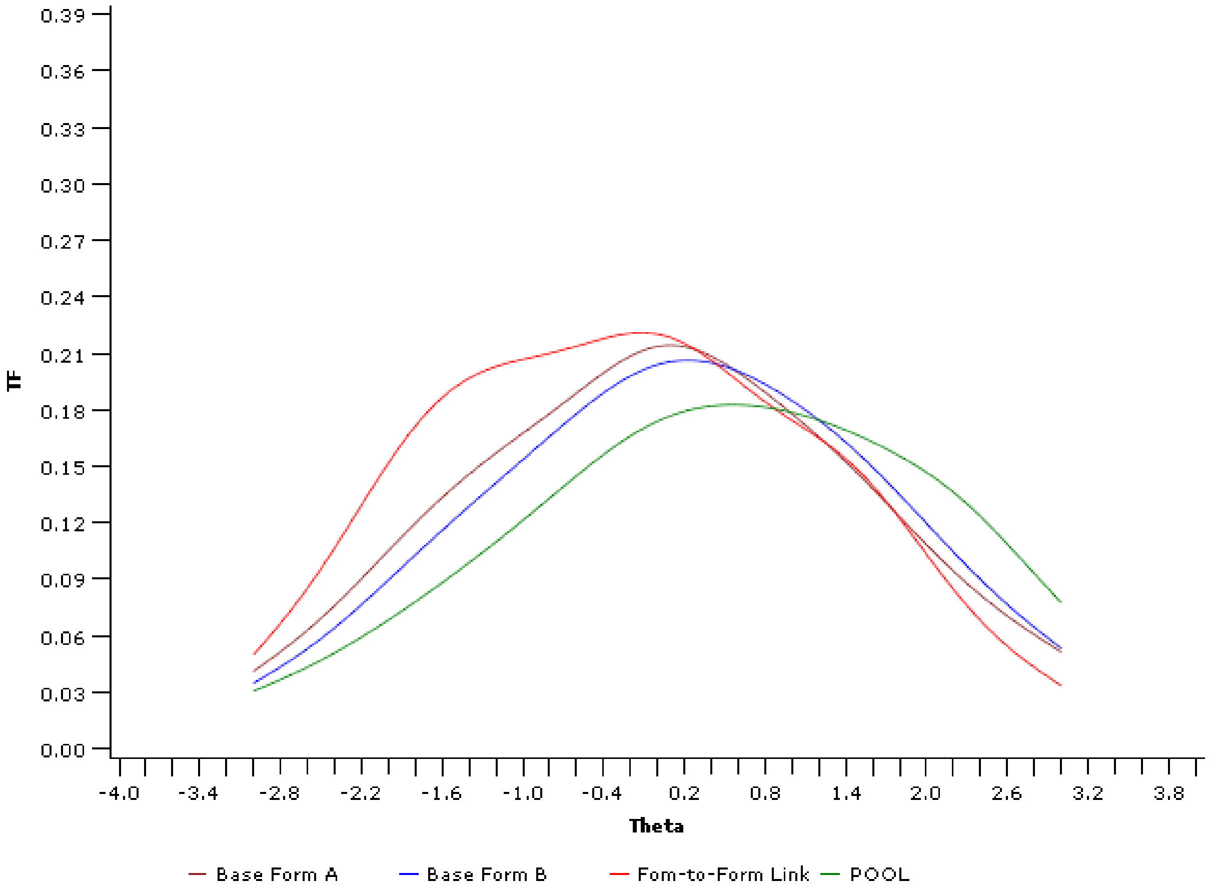
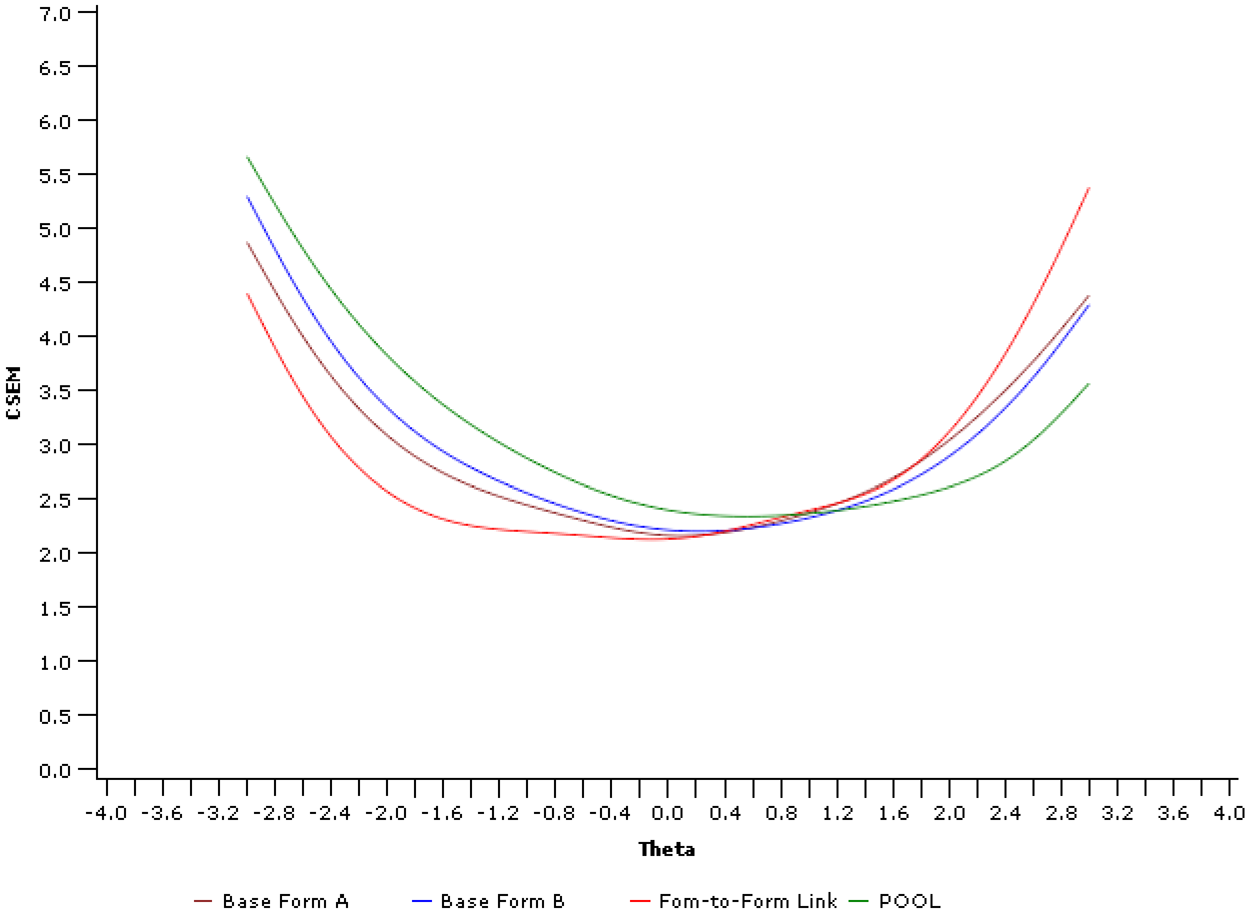


Figure 3. Conditional Standard Error of Measurement for the Grade 5 Science Test



Similar to grade 5, IRT item parameter estimates were used to generate characteristic curves (TCCs), test information functions (TIFs), and conditional standard errors of measure (CSEM) were computed for each of the base form, form-to-form linking items and entire item pool for grade 8. Figure 4 shows the overlaid TCC plots for Form A, B, linking item set and the entire item pool. The TCC and TIF values were divided by the total number of score points for each form so that the curves can be plotted on the same scale. Figure 5 displays test information curves for Form A, B, linking item set and the entire item pool. Figure 6 illustrates the conditional standard error of measurements for the four tests.

Figure 4. Test Characteristic Curve of the Grade 8 Science Test

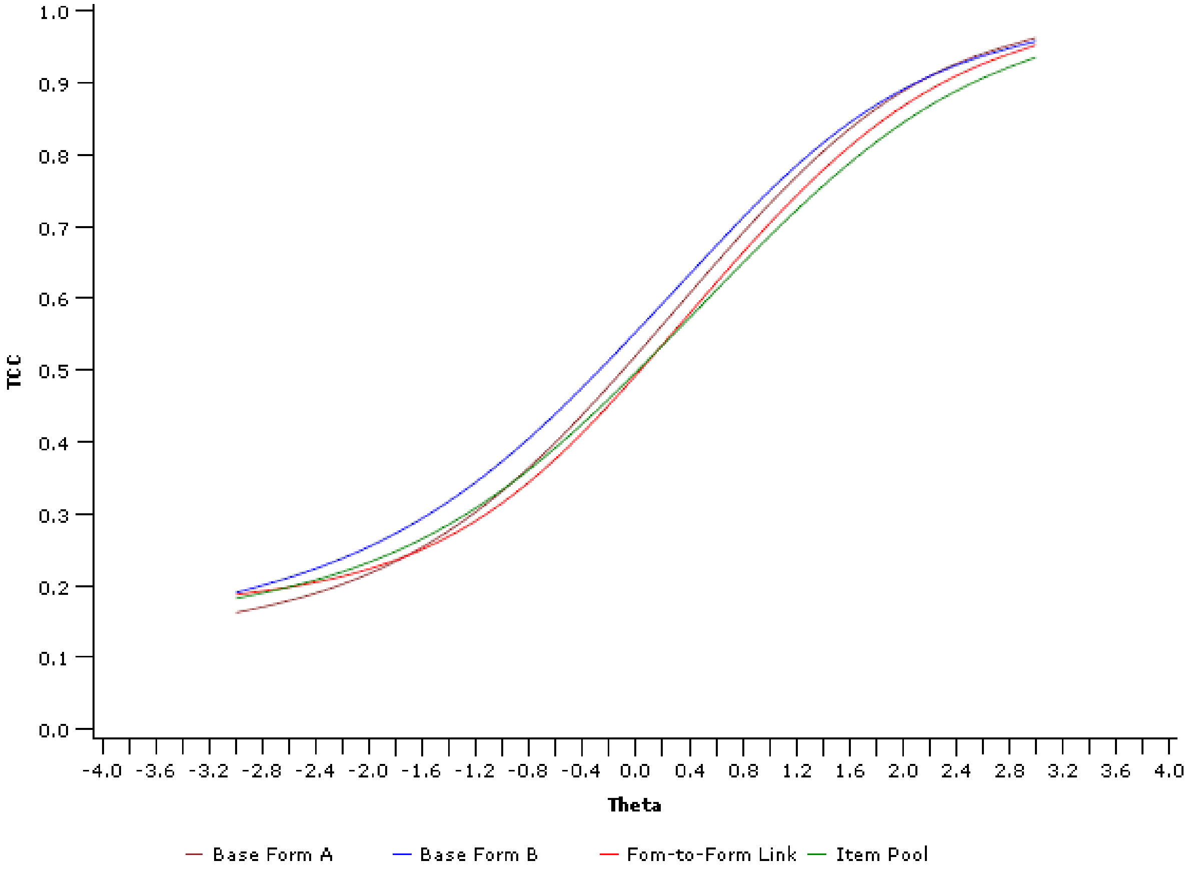


Figure 5. Test Information Function of the Grade 8 Science Test

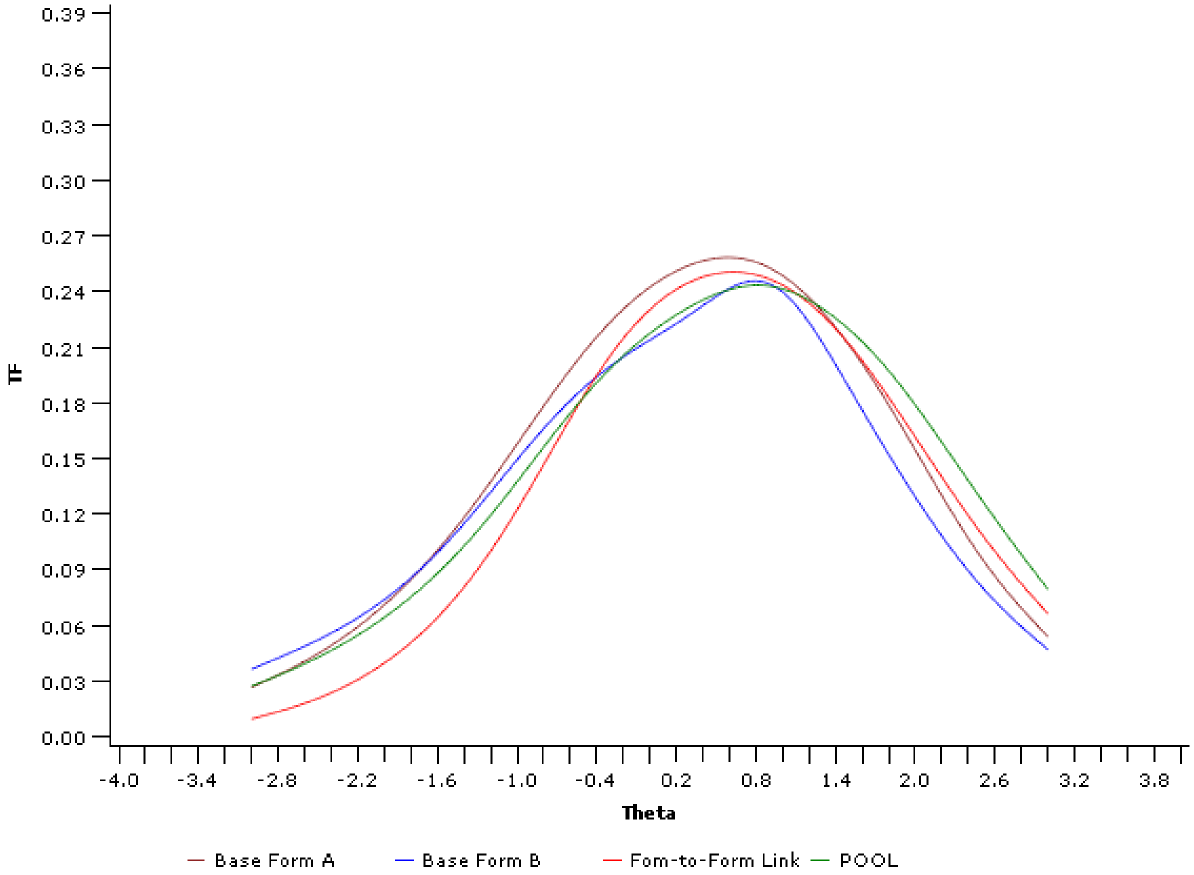
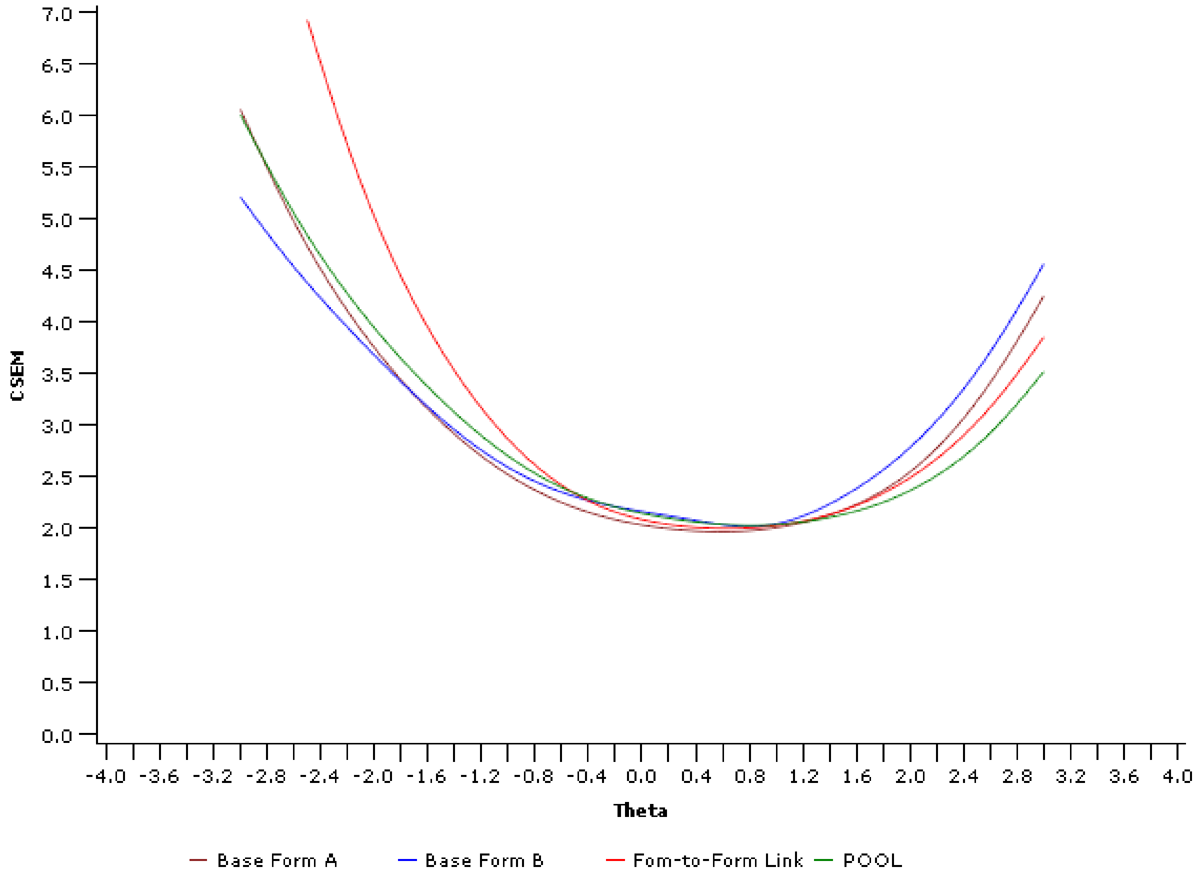


Figure 6. Conditional Standard Error of Measurement for Grade 8 Science Test



Defining Scale Ranges

In order to facilitate the use and interpretation of the results of the 2008 MSA Science operational administration, scale scores were created through the application of scaling constants developed following the 2007 test administration. Scale scores were computed using the following simple linear transformation equation:

$$SS = M1(\theta) + M2$$

where, M1 is a multiplicative term, M2 is an additive term, and θ is an IRT based measure of student ability. These scaling constants (M1 and M2) were developed to meet MSDE requirements that the mean and standard deviation (sd) be set at mean = 400 and sd = 40 on the scale score, while maintaining the LOSS at 240 and the HOSS at 650 as closely as possible for grades 5 and 8. The LOSS and HOSS set the minimum and maximum values that are possible on the MSA Science test. These scaling constants as well as the LOSS and HOSS for each grade appear in Table 8.

Table 8. Target LOSS, HOSS, and scaling constants for grades 5 and 8.

Grade	LOSS	HOSS	M1	M2
5	240	650	42.3077	400.1688
8	240	650	42.617	398.9311

In Maryland Science Assessment, student scale score was derived by item pattern scoring method based on maximum likelihood estimation. While maximum likelihood estimates were available for students with extreme scores other than zero or perfect, occasionally these estimates have very large conditional SEM (CSEM), and differences between these extreme values have little meaning. The ability estimates based on a relative small number of items as is the case for subscales tend to be unstable which can lead to a large CSEM. The CSEM for the extreme ability estimate therefore was truncated in consideration of the current MSA Science scale score range. The range helped us to maintain the student scores within the reasonable range while allowing us to have an understanding that a scale score of 240 is the lowest possible score a student can get on the test. As such, the range of CSEM should be maintained within a reasonable range.

Pearson proposed that a maximum SEM be set to be 160. The maximum SEM value is proposed based on multiple considerations.

- ***Relative magnitude of SEM to the scale score range.***
  Given the current scale score ranges from 240 to 650 which includes 410 points and the SEM is recommended not to exceed 40% of the scale score range. The SEM is an index to represent the measurement precision and the range in which the true student ability exists. A large SEM can lead to an interpretation that a student true ability can be either top or bottom of the scale. By curtailing the SEM to a reasonable value, we can provide a better estimate on where the student's true ability exists.
- ***Existing practice on other Maryland assessments***.
  According to the 2004 Maryland High School Assessment Technical Report, the SEM for LOSS and HOSS is set in consideration of the minimum SEM for the scale score. An internal and preliminary analysis on the Maryland Science SEM indicates that the minimum SEM for the scale score might be approximately 10 or 11 for grades 5 and 8.

Based on aforementioned considerations, the maximum CSEM was set to be 160. Upon the state approval of the recommendation, the truncation rule was implemented to report CSEM both for the overall score and the subscale scores.

ISE Pattern Scoring

In the spring 2008 administration of the MSA Science tests for grade 5 and 8, Pearson used an internally developed software program called IRT Score Estimation (ISE) program (Chien, Hsu, & Shin, 2007). The program has been extensively tested and compared to commercially available software programs (e.g., MULTILOG, PARSCALE; Tong, Um, Turhan, Parker, Shin, Chien, & Hsu, 2007). The report concluded that with normal cases the ISE program was able to replicate MULTILOG and PARSCALE theta estimates. However, "in problem cases, such as monotonically decreasing likelihood functions, in which MULTILOG and PARSCALE both produced theta estimates, ISE was able to produce the estimates that yielded the largest likelihood function, in alignment with the definition of the maximum likelihood algorithm" (p. 9). In addition, "with problem cases in which MULTILOG and PARSCALE failed to produce theta estimates, ISE was able to produce an estimate that yielded the largest likelihood from the likelihood function of a given response pattern" (p. 9). With regard to the CSEMs, ISE produced similar results to MULTILOG. More information about the ISE program can be found in the user manual, technical manual and evaluation report and are available upon request.

The 2008 operational scores were estimated by the pattern scoring approach. The 2008 operational item parameters were first equated to the base theta scale established in 2007. The equated item parameters were then used to estimate student ability (theta) using Pearson's ISE program. The theta estimates were transformed onto the MSA Science operational scale using the transformation constants described above.