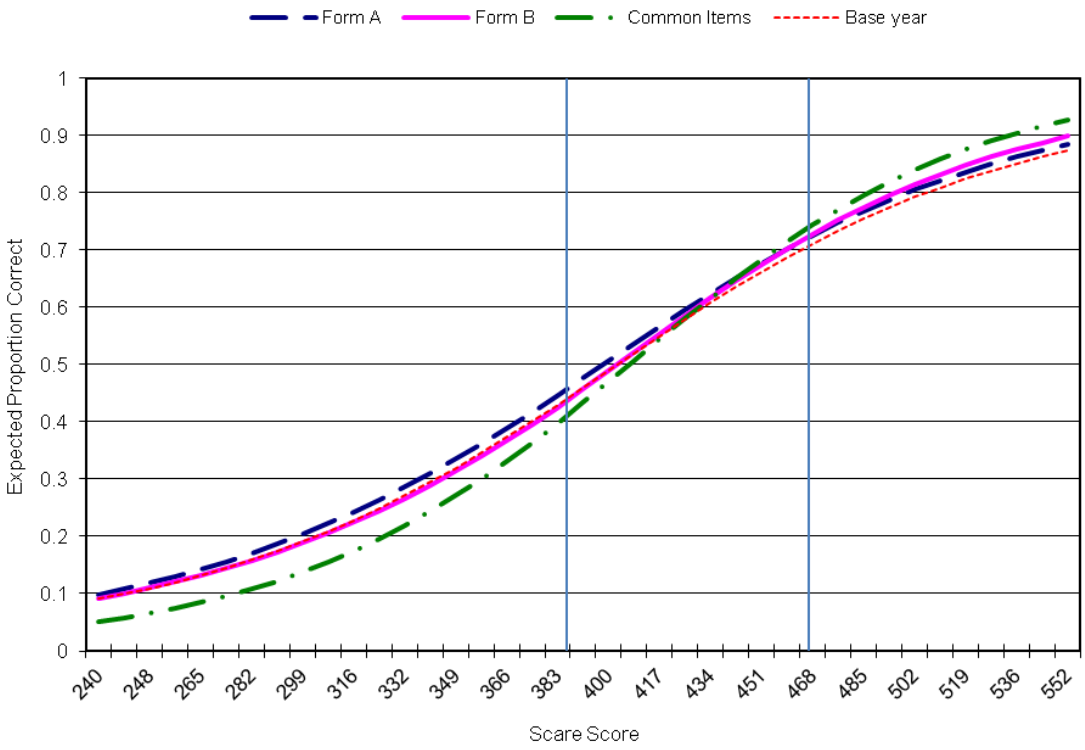
Test Analysis, Operational Scaling and Scoring

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 current year operational forms (A and B), form-to-form linking items (common items), and the base-year operational item pool. In order to facilitate comparisons of these curves, the TCC, TIF, and SEM values were divided by the total number of score points for each form so that the curves can be plotted on the same scale. These graphs show how well a given test form compares to another in terms of the measurement (scale) characteristics across the scale range. Here the primary comparisons are between the 2012 Form A and B curves and curves reflective of operational items from the 2008 (base year) administration.

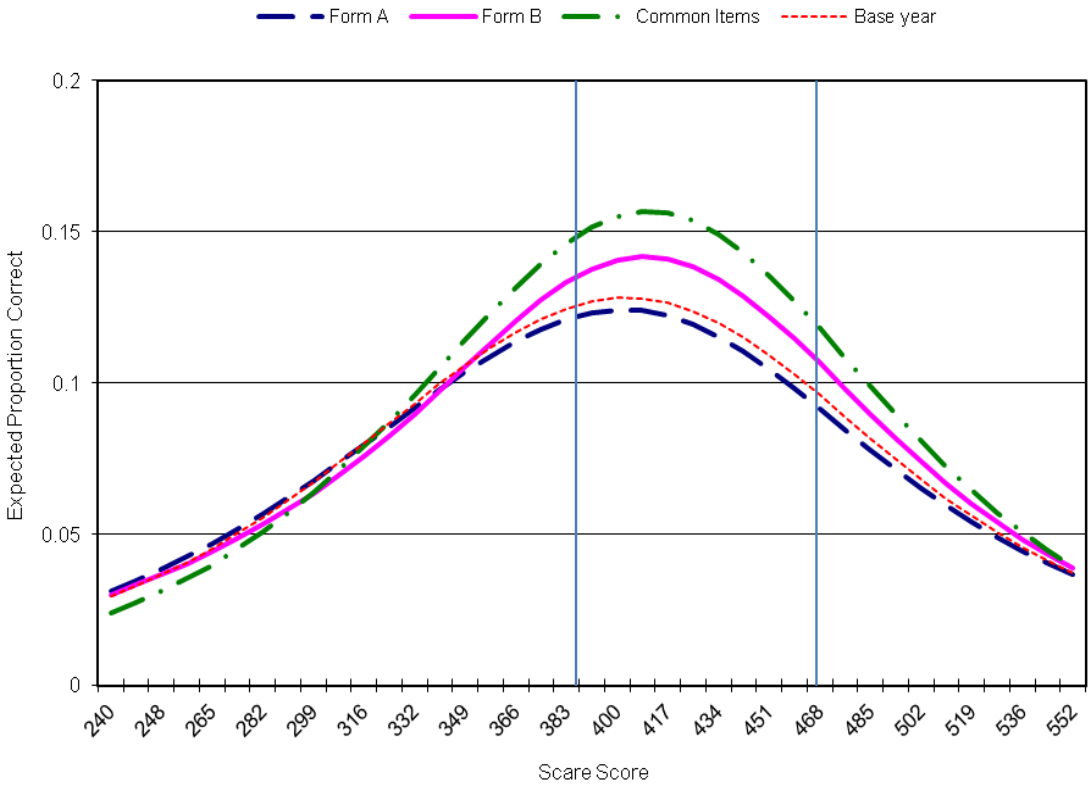
Figure 1 shows the overlaid TCC plots for Form A, Form B, form-to-form linking items and base-year item pool for grade 5. Figure 2 also displays test information curves for Form A, Form B, form-to-form linking items and the base-year. Figure 3 illustrates the conditional standard error of measurements for the four item sets. The vertical lines in each figure represent the location of the Proficient and Advanced performance standards on the reporting scale metric (each performance level is denoted at the top of the plot: Basic, Proficient, and Advanced). It should also be noted that each curve is presented according to the MSA Science scale score metric, which is described in the Defining Scale Ranges section.

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



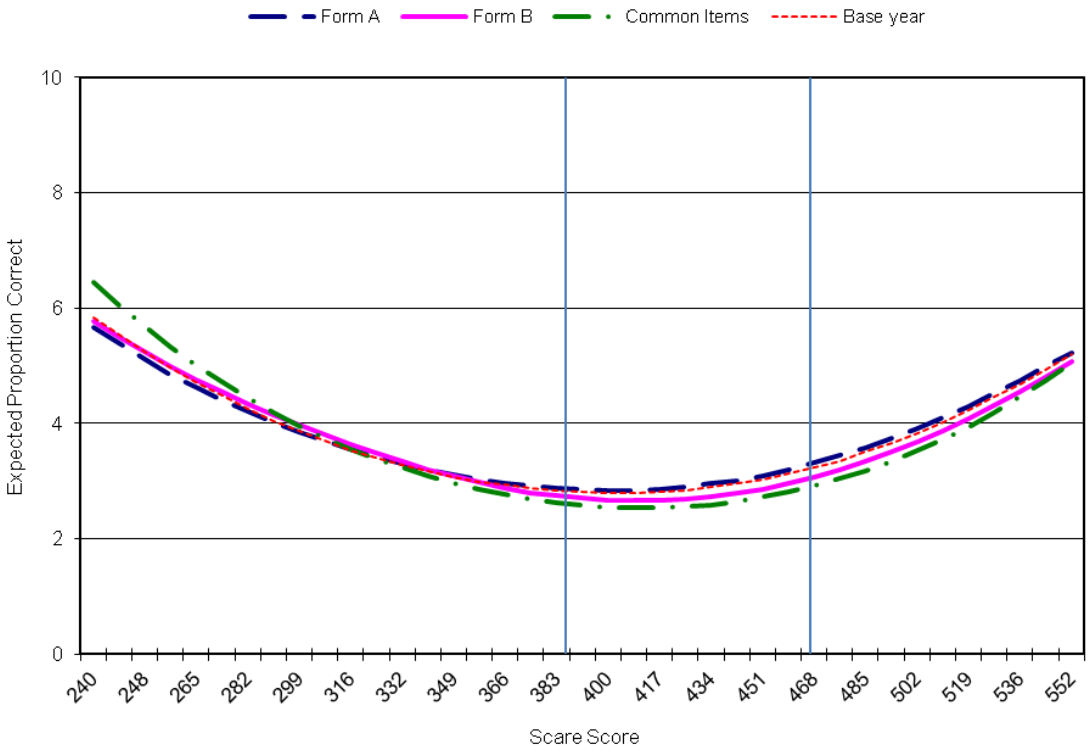
Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 391, Advanced Cut = 467).

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



Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 391, Advanced Cut = 467).

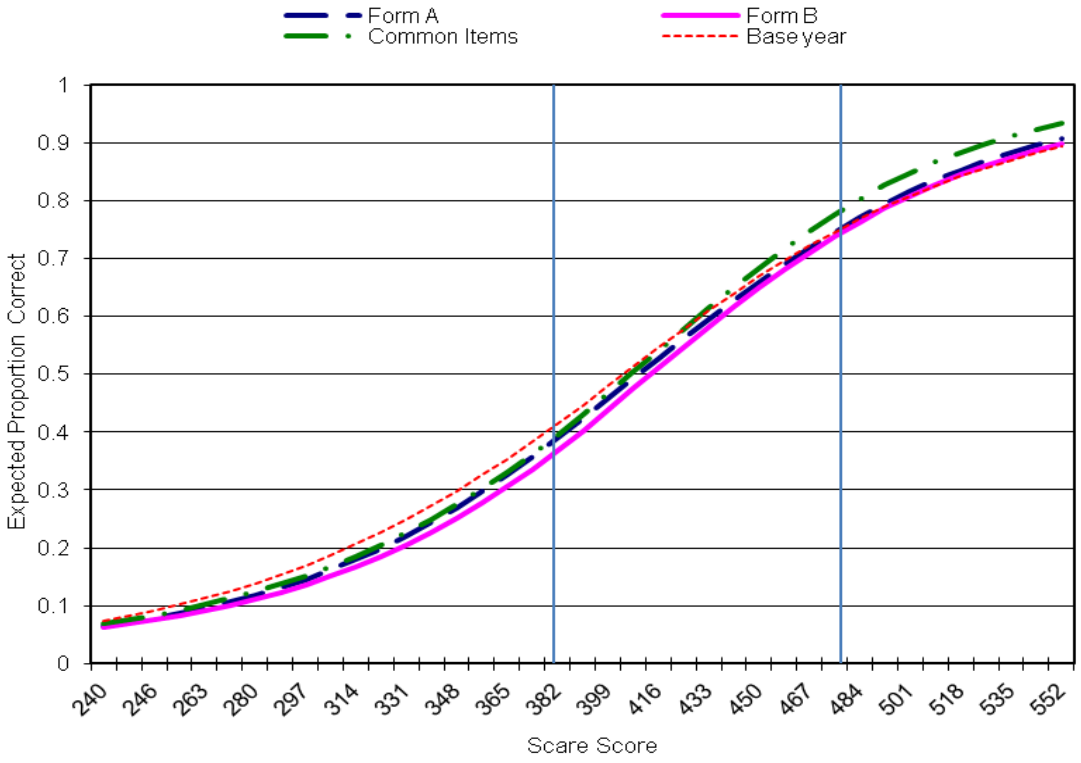
Figure 3. Conditional Standard Error of Measurement for the Grade 5 Science Test



Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 391, Advanced Cut = 467).

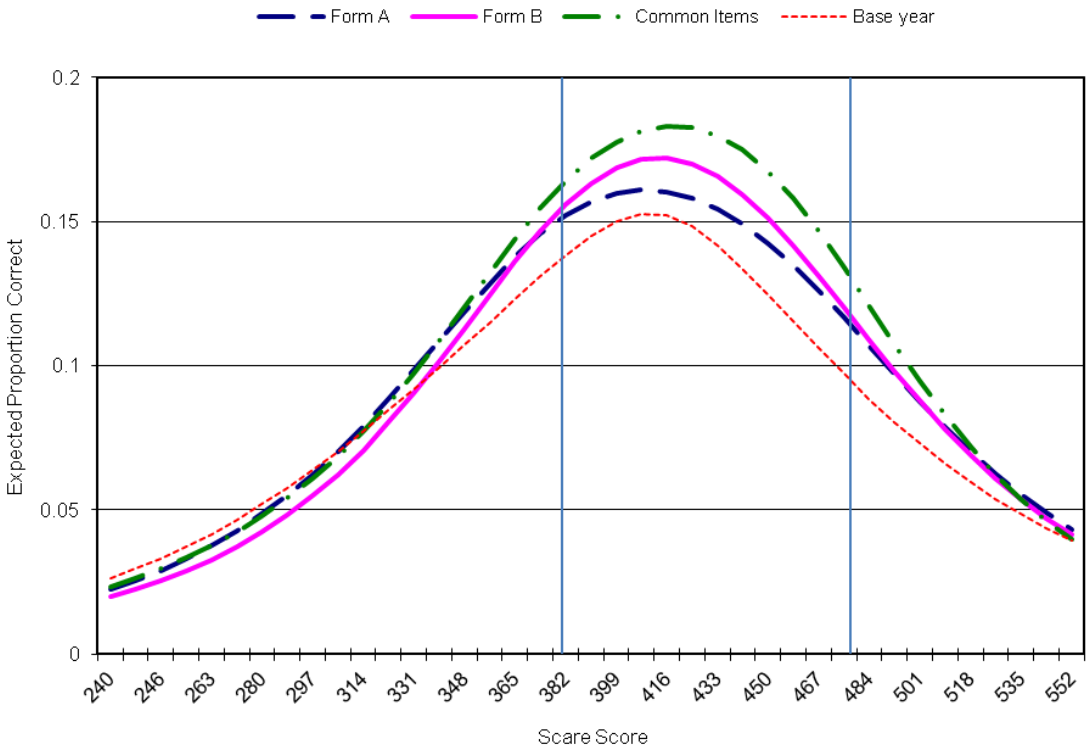
As with grade 5, IRT item parameter estimates were used to generate characteristic curves (TCCs), test information functions (TIFs), and conditional standard errors of measure (CSEM) were computed for each of the base forms, form-to-form linking items, and base-year operational test for grade 8. Figure 4 shows the overlaid TCC plots for Form A, B, linking item and base-year pools. 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 and base-year pools. Figure 6 illustrates the conditional standard error of measurements for the four item sets. The vertical lines in each figure represent the location of the Proficient and Advanced performance standards on the reporting scale metric. Note that each curve is presented relative to the scale score metric described in the Defining Scale Ranges section.

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



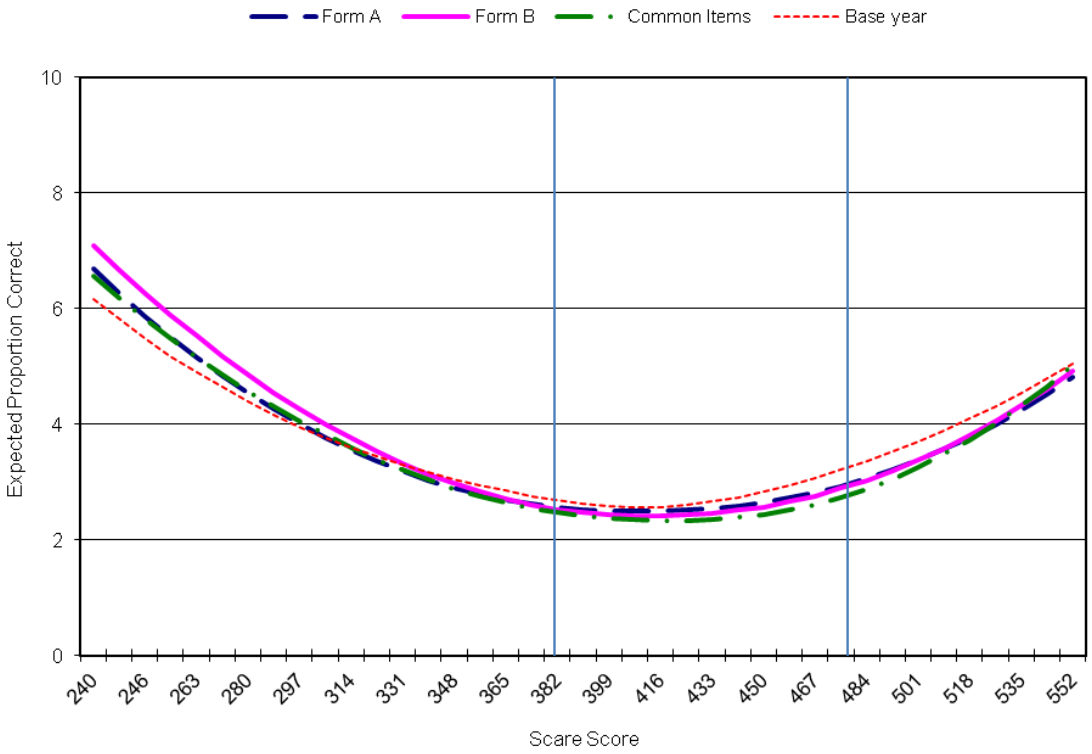
Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 387, Advanced Cut = 478).

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



Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 387, Advanced Cut = 478).

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



Note: The 2 vertical lines reflect the Proficient and Advanced cut scores which result in three performance levels: Basic, Proficient, and Advanced (Proficient Cut = 387, Advanced Cut = 478).

Defining Scale Ranges

The theta scale is not often used for reporting because of interpretation issues arising from a scale with values typically ranging from -4.0 to +4.0. Therefore, following the calibration and equating phases, the resulting theta values are transformed to a reporting scale that can be more meaningfully interpreted by students, teachers and other stakeholders. In order to facilitate the use and interpretation of the results of the 2012 MSA Science operational administration, scale scores were created through the application of scaling constants determined from the base 2007 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 established in the base year at mean scale score = 400 and sd = 40, while maintaining the lowest obtainable scale score (LOSS) at 240 and the highest obtainable scale score (HOSS) at 650. 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 7.

Table 7. 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

ISE Pattern Scoring

Pearson used an internally developed software program called IRT Score Estimation (ISE; Chien, Hsu, & Shin, 2007) to conduct pattern scoring for the spring 2012 administration of the MSA Science tests for grades 5 and 8. 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 CSEM, ISE produced similar results to MULTILOG. More information about the ISE program can be found in the user manual, the technical manual, and the evaluation report, which are available upon request.

The 2012 operational scores were estimated by the pattern scoring approach. The 2012 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. It should be noted that one SR item in grade 5 was not used for equating or scoring purposes because it had been previously released and overall impact was negligible. Final theta estimates from ISE were transformed onto the MSA Science operational scale using the scaling constants described above.

Conditional Standard Errors for LOSS and HOSS

Within ISE, student ability (theta) is determined via maximum likelihood estimation (MLE). One characteristic of MLE is that for students with scores of zero or perfect scores, abilities are not estimable (i.e., they effectively result in estimates of $\pm \infty$). Because of this it is typical to establish ability values or scale scores that are in line with the respective overall scale. For the MSA Science tests, the LOSS and HOSS values reflect the values associated with these extreme scores. Additionally, there are instances in which certain score patterns close to zero and perfect scores will provide ability estimates where the respective conditional standard errors of measurement (CSEM) are very large. These inflated CSEM estimates are problematic in that they are out of line with estimates from different score patterns but of the same ability. In addition to establishing reasonable scale scores for these points, it is also desirable to provide some reasonable associated standard error to promote appropriate score interpretation.

In order to provide students with appropriate score interpretations where ability estimates from the MSA Science tests are associated with the LOSS and HOSS scale scores (240 and 650), and Pearson recommended a maximum CSEM of 160 be used. This recommendation was based on multiple considerations.

First of all, consideration was given to the magnitude of standard errors relative to the overall scale score range. The current scale ranges from 240 to 650 (410 total points). When standard

errors exceed 40% of a scale range, the utility of a test score interpretation is limited. With this in mind, the initial 2007 MSA Science base scaling was evaluated.

The initial 2007 MSA Science administration involved the administration of ten field test forms per grade; each created in line with the MSA Science blueprints and served as the mechanism for establishing the base scales. For each form, ability estimates were generated and their associated standard errors were examined. Across grade 5 and 8 forms, the largest standard errors for the highest estimable abilities were roughly 155 scale score points and were within the 40% heuristic noted above.

In addition to evaluation of the base year calibrations, consideration was also given to standing practice for other Maryland assessments; specifically the Maryland High School Assessments (HSA). The 2004 HSA Technical Report describes principals adopted for the determination of optimal LOSS and HOSS values where associated standard errors are also described (Appendix 3.C). In determining a value for HOSS, it was recommended that the associated conditional standard error be lower than ten times the minimum conditional standard error on the overall test. For the LOSS, the recommendation was for the associated conditional standard error to be lower than fifteen times the minimum conditional standard error on the test. For the base year MSA Science administration, minimum CSEM values were roughly 11 scale score points.

Based on these considerations, a recommendation was made for the maximum CSEM be set to 160 for the LOSS and HOSS. This was in line with the observed standard errors from the base year calibrations for extreme scores and also in line with existing practice. Upon state approval of the recommendation, the rule was implemented to report CSEM for all scores.

Test Score Reliability

The reliability of a test provides an estimate of the extent to which an assessment will yield the same results across subsequent administrations, provided the two administrations do not differ on relevant variables. Reliability coefficients are usually forms of correlation coefficients and must be interpreted within the context and design of the assessment and of the reliability study. The forms of reliability below measure different dimensions of reliability and thus any or all might be used in assessing the reliability of MSA Science.

The estimates of reliability reported here are measures of internal consistency and reflect the degree to which the components of a test are consistent with other components of the test. One of the most commonly used indices of internal consistency reliability is Cronbach's coefficient *alpha* (α; Cronbach, 1951). In this formula, the s_i^2 denotes the variances for the k individual items; s_{sum}^2 denotes the variance for the sum of all items.

$$\alpha = (k/(k-1)) * [1- \Sigma(s^2_i)/s^2_{sum}]$$

Because of the mixed item types on the MSA Science test (i.e., SR and BCR), a stratified alpha (Cronbach, Schönemann, & McKie, 1965) is more appropriate. Stratified alpha accounts for the fact that different groups of items ("strata") may have different variances. Since the Cronbach alpha relies on a single overall variance, it may not be the best estimate of "true" reliability. Because of this, stratified alpha reliability coefficients were computed for the MSA Science tests. The formula is:

$$Stratified\ \alpha = 1 - \frac{((\sigma_{SR}^2(1-\rho_{SR}) + (\sigma_{CR}^2(1-\rho_{CR}))}{\sigma_t^2}$$

where

σ_{SR}^2 = variance associated with SR items;

σ_{CR}^2 = variance associated with BCR items;

σ_t^2 = variance of total score;

ρ_{SR} = reliability associated with the SR items; and

ρ_{CR} = reliability associated with BCR items.

These results are presented in Table 8.

Table 8. Reliability Estimate by Grade, Form, Gender and Ethnicity

Group		Grade 5		Grade 8	
		Form A	**Form B**	**Form A**	**Form B**
Overall		0.92	0.92	0.93	0.93
Gender	Female	0.92	0.92	0.92	0.93
	Male	0.92	0.93	0.93	0.94
Ethnicity	Hispanic/ Latino	0.91	0.91	0.92	0.92
	Non-Hispanic/ Latino	0.92	0.92	0.93	0.93
Race	African American	0.90	0.90	0.90	0.91
	American Indian	0.91	0.92	0.91	0.92
	Asian/Pacific Islander	0.91	0.92	0.92	0.94
	Native Hawaiian	0.97	0.94	0.91	0.93
	White	0.90	0.91	0.91	0.92

The coefficient alpha estimates for all forms meet conventional guidelines for applied test reliability (i.e., α > .85).