Section 3. Scoring Procedures and Score Types

Scale Scores

Scale scores based on maximum likelihood estimates (MLE) were reported for the total test score. All scores were reported on the operational reporting scale established in 2003. While the total test score was based on item-pattern (IP) scoring, the subscores were based on number-correct (NC) to scale score scoring tables.

With IP scoring, because the likelihood equation can have multiple maxima with the 3PL model, a numerical method was developed that found the scale score at the global maximum in the likelihood function. NC to scale score scoring tables were obtained by inversing the test characteristic curves (TCC) of items contributing to the associated subscores and this procedure produced what Yen (1984) called 'number correct trait estimates'. In this report, we call it 'NC scale scores'.

Prior to commencing with the 2004 scoring, MSDE had asked ETS to investigate and replicate the 2003 analyses for the English High School test completed by their previous vendor, CTB/McGraw-Hill. Using independent software, we were able to replicate the results, although small differences were noted in the parameter estimates, transformation constants, and mean scores. However, this is to be expected due to variations associated with inclusion/exclusion criteria for the calibration sample, and differences in the calibration software. Based on the results of this study, we also found no evidence of a systematic error or problem with the calibrations and linking studies completed by CTB/McGraw-Hill. The complete results of the study are presented in Appendix 3.A.

Conditional Standard Errors of Measurement.

Corresponding conditional standard errors of measurement (SEM) were also produced for both types of scoring and were equal to the inverse of the square root of the test information function.

$$\mathrm{SEM}(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$$

where,

SEM($\hat{\theta}$)=standard error of measurement

$I(\theta)$= test information function.

The test information function is the sum of corresponding information functions of the test items when optimal item weights are used, as in the HSAs. Item information functions depend on the item difficulty, discrimination and conditional item score variance. Thus, while polytomous items often have lower discriminations than selected response items (Fitzpatrick et al., 1996) they may convey more information than selected response items, because they have more score points.

The SEM curves for each test were presented in Section 1 (see Figures 1.2 for Algebra, Figure 1.4 for Biology, Figure 1.6 for English I, Figure 1.8 for Geometry and Figure 1.10 for Government). As can be observed in these figures, the SEMs vary across the scale. In all cases, extreme values were noted at the ends of the scale, but the SEM is minimized near the cut-scores for each content area, which were near the middle of the scale. This pattern is expected as 1) more items tend to be of middle difficulty; and 2) there were fewer items at the lower and upper ends of the scale. In all cases the SEM is less than 10 scale score points at the cut point.

Subscore Scoring

For the subscore scale scores, the NC to scale score scoring method (later called the NC scoring) was selected based on a special study that compared the two different scoring methods (see Appendix 3.B). At the classroom level, which is where these scores were used, the IP and NC methods produced nearly identical means for all subscores except the one with the fewest score points. This is consistent with other studies that have identified that while IP and NC ability estimates differ for individual examinees (i.e., for examinees with the same number-correct score, their item-pattern ability estimate may be higher or lower, depending on which items they got correct), these two ability estimates were tau-equivalent for groups of 30 or more examinees (Yen, 1984; Yen & Candell, 1991). While the benefit of using IP scoring is the reduced conditional SEMs relative to NC scoring, for the subscore with the fewest score points, IP scores had much higher conditional SEMs than NC scores through the lower part of the score scale. This occurred because a much larger number of scores were assigned the LOSS using IP scoring compared to NC scoring. The difference in results was caused by differential "interpretation" by the IP and NC methods of low scores that did/did not include score points earned on constructed response items. Essentially, IP scoring was not observed to be uniformly beneficial for subscores when there were a small number of score points that included both SR and CR items, and for subscores, the NC scoring method was subsequently recommended by the National Psychometric Committee (NPC).

Lowest and Highest Obtainable Test Scores

Both maximum likelihood procedure and NC scoring cannot produce scale score estimates for students with perfect scores or scores below the level expected by guessing. Also, while maximum likelihood estimates were available for students with extreme scores other than zero or perfect, occasionally these estimates have very large conditional SEMs, and differences between these extreme values have little meaning. Therefore, scores were established for these students based on a rational procedure (see Appendix 3.B; CTB/McGraw-Hill, December 2003). These values were called the lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS). The same LOSS and HOSS values were used for either number-correct (NC) or item-pattern scoring. In addition, the associated conditional SEMs were constrained to a maximum value of 80. Table 3.1 lists the LOSS and HOSS scores for each content area established following the first operational administration (CTB/McGraw-Hill, December, 2003).

Table 3.1 LOSS and HOSS Values

Content	LOSS	HOSS
Algebra	200	625
Biology	225	650
English I	200	625
Geometry	225	600
Government	225	650

Cut-Scores

The cut-scores associated with each of the performance levels in each of the content areas were established by MSDE in 2003 (see Table 3.2). One cut-score was established for all of the content areas except for Geometry. Because Geometry is used as the high school mathematics component of the MD accountability plan under NCLB, two cut-scores were established.

Table 3.2 HSA 2004 Cut-Scores

Content Area	Cut-score
Algebra	412
Biology	400
English I	407
Government	394
Geometry	Proficient – 411
	Advanced – 447