Section 5. Field Test Analyses

Following the receipt of the final scored file from Measurement Incorporated (MI), the field test analyses were completed. The analyses of the field test data consisted of four components: classical item analyses, differential item functioning (DIF), calibration, and scaling. All of the analyses were completed using Genasys, ETS proprietary software. The analysis procedures for each component are described in detail. Samples used for the analyses included all valid records available at the time of the analyses, including students classified as English as a second language, students with IEP or 504 plans, and students receiving accommodations. Only duplicate records, records invalidated by the test administrator, and records with five or fewer item responses were excluded from the analysis sample. The field test analyses presented in this section reflect only the January 2005 administrations. The May 2005 field test data were not available when the draft report was prepared. The May 2005 field test analyses will be presented in the final version of this report.

Classical Item Analyses

Classical item analyses involve computing a set of statistics based on classical test theory for every item in each form. The statistics provide key information about the quality of the items from an empirical perspective. The statistics estimated for the MHSA field test items are described below.

Classical item difficulty ("P-Value"):

This statistic indicates the percent of examinees in the sample that answered the item correctly. Desired p-values generally fall within the range of 0.25 to 0.90. Occasionally, items that fall outside this range can be justified for inclusion in an item bank based upon the quality and educational importance of the item content or the ability to measure students with very high or low achievement, especially if the students have not yet received instruction in the content or lack motivation to complete the field test items to the best of their ability.

The item-total correlation of the correct response option (for SR items) or the CR item score with the total test score:

This statistic describes the relationship between performance on the specific item and performance on the entire form. It is sometimes referred to as a discrimination index. Values less than 0.15 were flagged for a weaker than desired relationship and deserve careful consideration by ETS staff and MSDE before including them on future forms. Items with negative correlations can indicate serious problems with the item content (e.g., multiple correct answers, unusually complex content), an incorrect key, or students have not been taught the content.

The proportion of students choosing each response option (SR items):

This statistic indicates the percent of examinees selecting each answer option. Item options not selected by any students or selected by a very low proportion of students indicate problems with plausibility of the option. Items that do not have all answer options functioning may be discarded or revised and field tested again.

The point-biserial correlation of incorrect response option (SR items) with the total score:

These statistics describe the relationship between selecting an incorrect response option for a specific item and performance on the entire test. Typically, the correlation between an incorrect answer and total test performance is weak or negative. Values are typically compared and contrasted with the discrimination index. When the magnitude of these point-biserial correlations for the incorrect answer is stronger, relative to the correct answer, the item will be carefully reviewed for content-related problems. Alternatively, positive point-biserial correlations on incorrect option choices may indicate that students have not had sufficient opportunity to learn the material.

Percent of students omitting an item:

This statistic is useful for identifying problems with test features such as testing time and item/test layout. Typically, we would expect that if students have an adequate amount of testing time, 95% of students should attempt to answer each question. When a pattern of omit percentages exceeds 5% for a series of items at the end of a timed section, this may indicate that there was insufficient time for students to complete all items. Alternatively, if the omit percentage is greater than 5% for a single item, this could be an indication of an item/test layout problem. For example, students might accidentally skip an item that follows a lengthy stem.

Frequency distribution of CR score points:

Observation of the distribution of scores is useful to identify how well the item is functioning. If no students are assigned the top score point, this may indicate that the item is not functioning with respect to the rubric, there are problems with the item content, or students have not been taught the content.

Summaries of p-values by content area for the field test items administered in January are found in Table 5.1 for SR items and Table 5.2 for CR items. Summaries of item-total correlations by content area for the field test items administered in January are found in Table 5.3 for the SR items and Table 5.4 for the CR items. In addition, a series of flags was created to identify items with extreme values. Flagged items were subject to additional scrutiny prior to the inclusion of the items in the final calibrations. The following flagging criteria were applied to all items tested in the 2005 assessments:

- *Difficulty Flag*: P-values less than 0.25 or greater than 0.90.
- *Discrimination Flag*: Point-biserial correlation less than 0.15 for the correct answer.
- *Distractor Flag*: Point-biserial correlation positive for incorrect option.

- *Omit Flag*: Percentage omitted is greater than 0.05.
- *Collapsed Score Levels*: Items with no students obtaining the score point.

Following the classical item analyses, items with poor item statistics and items that were not scored were removed from further analyses. Refer to Table 5.5. These items have been identified for revision and possible re-field testing.

Differential Item Functioning (DIF)

Following the classical item analyses, DIF analyses were completed. One goal of test development is to assemble a set of items that provides an estimate of a student's ability that is as fair and accurate as possible for all groups within the population. DIF statistics are used to identify items whereby identifiable groups of students with the same underlying level of ability have different probabilities of answering correctly (e.g. females, African Americans, Hispanics). If the item is more difficult for an identifiable subgroup, the item may be measuring something different than the intended construct. However, it is important to recognize that DIF flagged items might be related to actual differences in relevant knowledge or skill (item impact) or statistical Type I error. As a result, DIF statistics are used to identify potential sources of item bias. Subsequent review by content experts and bias/sensitivity committees is required to determine the source and meaning of evident differences.

ETS used two DIF detection methods: the Mantel-Haenszel and standardization approaches. As part of the Mantel-Haenszel procedure, the statistic described by Holland & Thayer (1988), known as MH D-DIF, was used[2]. This statistic is expressed as the differences between the focal and reference group after conditioning on total test score. The statistic is reported on the ETS delta scale, which is a normalized transformation of item difficulty (proportion correct) with a mean of 12 and a standard deviation of 4.

[2] The formula for the estimate of constant odds ratio is:

$$\alpha_{MH} = \frac{\left(\sum_m \frac{R_{rm} W_{fm}}{N_m}\right)}{\left(\sum_m \frac{R_{fm} W_{rm}}{N_m}\right)},$$

where,

R_{rm} = number in reference group at ability level m answering the item right,
W_{fm} = number in focal group at ability level m, answering the item wrong,
R_{fm} = number in focal group at ability level m answering the item right,
W_{rm} = number in reference group at ability level m, answering the item wrong,
N_m = total group at ability level m.

This can then be used in the following formula (Holland & Thayer, 1985):
$MH\ D\text{-}DIF = -2.35\ ln[\alpha_{MH}]$.

Negative MH D-DIF statistics favor the reference group and positive values favor the focal group. The classification logic used for flagging items is based on a combination of absolute differences and significance testing. Items that are not statistically significantly different based on the MH D-DIF ($p>0.05$) are considered to have similar performance between the two studied groups; these items are considered to be functioning appropriately. For items where the statistical test indicates significant differences ($p < 0.05$), the effect size is used to determine the direction and severity of the DIF. For the ELA CR item, the Mantel-Haenszel procedure was executed where item categories are treated as integer scores and a chi-square test was carried out with one degree of freedom. The male and white groups were considered the reference groups for gender and ethnicity, respectively; the female and other ethnic groups were considered the focal groups.

Based on these DIF statistics, items are classified into one of three categories and assigned values of A, B or C. Category A items contain negligible DIF, Category B items exhibit slight or moderate DIF, and Category C items have moderate to large DIF. Negative values imply that conditional on the matching variable, the focal group has a lower mean item score than the reference group. In contrast a positive value implies that, conditional on the matching variable, the reference group has lower mean item score than the focal group. For constructed-response items the MH D-DIF is not calculated, but analogous flagged rules based on the chi-square statistic have been developed resulting in classification into A, B, or C DIF categories.

There were 8 items flagged for C-level DIF against one of the identified focal groups (i.e., female, African American, American Indian, Asian, Hispanic) for two of the four content areas. For the government test, 2 items were flagged to have negative DIF against female (favor female), 3 items against African American (favor White) and 1 item against Hispanic (favor White). For the Biology test, 1 item was flagged to have DIF against African American (favor White) and another item was flagged to have negative DIF against African American (favor African American). These items are flagged in the bank, and will be reviewed for future use.

IRT Calibration and Scaling

The purpose of item calibration and scaling is to create a common scale for expressing the difficulty estimates of all the items across all versions of a test. The resulting scale has a mean score of 0 and a standard deviation of 1. It should be noted that this scale is often referred to as the "theta" metric and is not used for reporting purposes because the values typically range from –3 to +3. Therefore, the scale is usually transformed to a reporting scale (also know as a scale score), which can be more meaningfully interpreted by students, teachers, and other stakeholders.

The IRT models used to calibrate the MHSA test items were the 3-parameter logistic (3PL) model for SR items and the generalized partial credit model (GPCM) for CR items. Item response theory expresses the probability that a student will achieve a certain score

on an item (such as correct or incorrect) as a function of the item's statistical properties and the ability level (or proficiency level) of the student.

The fundamental equation of the 3PL model relates the probability that a person with ability θ will respond correctly to item j:

$$P(U_j = 1 \mid \theta) = P_j(\theta) = c_j + \frac{1 - c_j}{1 + e^{-1.7a_j(\theta - b_j)}}$$

where:

U_j is the response to item j, 1 if correct and 0 if incorrect;

a_j is the slope parameter of item j, characterizing its discriminating power;

b_j is the threshold parameter of item j, characterizing its difficulty; and

c_j is the lower asymptote parameter of item j, reflecting the chance that students with very low proficiency will select the correct answer, sometimes called the "pseudo-guessing" level

The parameters estimated for the 3-PL model were discrimination (a), difficulty (b), and the pseudo-guessing level (c).

The GPCM is given by

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{v=1}^{k} Z_{jv}(\theta)\right]}{\sum_{c=1}^{m_j} \exp\left[\sum_{v=1}^{c} Z_{jv}(\theta)\right]}$$

where

$$Z_{jk}(\theta) = 1.7a_j(\theta - b_{jk}) = 1.7a_j(\theta - b_j + d_k)$$

$$\sum_{k=2}^{m_j} d_k = 0$$

P_{jk} is the probability of responding in the k^{th} category from m_j+1 categories for item j,

θ is the ability level,

a_j is the item parameter characterizing the discriminating power for item j,

b_{jk} is an item-category parameter for item j,

b_j is the item parameter characterizing the difficulty for item j,

d_k is the category parameter characterizing the relative difficulty of category k .

A proprietary version of the PARSCALE computer program (Muraki & Bock, 1995) was used for all item calibration work. This program estimates parameters for a generalized partial-credit model using procedures described by Muraki (1992). The resulting calibrations were then scaled to the bank estimates using the Stocking and Lord's (1983) test characteristic curve method using the operational items as the "anchor" set.

The calibration and equating process is outlined in the steps below:

1. For each test, calibrate all items using a sparse matrix design that places all items on a common scale. Essentially, this means that the data was analyzed using the following format. In the diagram below X's represent items, spaces indicating missing data. For example, items included on version 2 but not on version 1, 3, 4 or 5 were treated as "not reached" for the purposes of the analyses and were denoted as "missing" in the diagram below.

```
Common    Unique 1       Unique 2          Unique 3         Unique 4         Unique 5
XXXXXXXXXXXXXXXX
XXXXXX              XXXXXXXXXX
XXXXXX                        XXXXXXXXX
XXXXXX                                 XXXXXXXXX
XXXXXX                                           XXXXXXXXX
```

2. Once the items have been calibrated, results are reviewed to determine if any items failed to calibrate.

3. After the final calibration parameters were obtained, the items were then linked to the bank scale using the test characteristic curve method. Specifically, the operational items were used to place the field test items onto the operational reporting scale.

Once the items were calibrated and placed onto the operational scale, the items were loaded into the item bank. Items were listed as unavailable based on the following criteria:

- Item-total correlation less than 0.1
- Item P-value less than 0.1
- Field test CR items that have fewer than 20 students achieving the highest score level
- Item not scored

Statistical Summary Tables

Table 5.1. Distribution of P-Values for the January Field Test SR Items

	Percentage and Number of Items							
P-Value	Algebra		Biology		Geometry		Government	
	%	N	%	N	%	N	%	N
< 0.30	12.5	2	11.1	3	13.3	2	4.2	1
0.30 to 0.40	43.8	7	14.8	4	6.7	1	12.5	3
0.41 to 0.50	12.5	2	25.9	7	26.7	4	33.3	8
0.51 to 0.60	0	0	18.5	5	20.0	3	12.5	3
0.61 to 0.70	18.8	3	18.5	5	33.3	5	16.7	4
0.71 to 0.80	12.5	2	7.4	2	0	0	20.8	5
≥ 0.81	0	0	3.7	1	0	0	0	0
Descriptive Stats								
Number of Items	16		27		15		24	
Mean	0.44		0.51		0.50		0.54	
SD	0.18		0.16		0.17		0.17	
Min	0.11		0.22		0.12		0.26	
Max	0.75		0.91		0.70		0.80	

Table 5.2. Distribution of P-Values for the January Field Test CR Items

	Percentage of Items (N)							
P-Value	Algebra		Biology		Geometry		Government	
	%	N	%	N	%	N	%	N
< 0.30	50.0	2	100	1	50.0	2	75.0	3
0.30 to 0.40	0	0	0	0	25.0	1	25.0	1
0.41 to 0.50	50.0	2	0	0	25.0	1	0	0
0.51 to 0.60	0	0	0	0	0	0	0	0
0.61 to 0.70	0	0	0	0	0	0	0	0
0.71 to 0.80	0	0	0	0	0	0	0	0
≥ 0.81	0	0	0	0	0	0	0	0
Descriptive Stats								
Number of Items	4		1		4		4	
Mean	0.30		0.17		0.29		0.20	
SD	0.16		0.00		0.10		0.11	
Min	0.11		0.17		0.19		0.10	
Max	0.44		0.17		0.41		0.33	

Table 5.3 Distribution of Item-Total Correlations for the January Field Test SR Items

	Percentage and Number of Items							
Correlation	Algebra		Biology		Geometry		Government	
	%	N	%	N	%	N	%	N
< 0.15	0	0	7.4	2	0	0	12.5	3
0.15 to 0.24	18.8	3	14.8	4	6.7	1	16.7	4
0.25 to 0.34	37.5	6	22.2	6	6.7	1	8.3	2
0.35 to 0.44	18.8	3	37.0	10	46.7	7	50.0	12
0.45 to 0.54	25.0	4	18.5	5	20.0	3	12.5	3
≥ 0.55	0	0	0	0	20.0	3	0	0
Descriptive Stats								
Number of Items	16		27		15		24	
Mean	0.35		0.34		0.45		0.33	
SD	0.11		0.11		0.12		0.11	
Min	0.20		0.13		0.24		0.12	
Max	0.54		0.52		0.73		0.49	

Table 5.4 Distribution of Item-Total Correlations for January Field Test CR Items

	Percentage and Number of Items							
Correlation	Algebra		Biology		Geometry		Government	
	%	N	%	N	%	N	%	N
< 0.15	0	0	0	0	0	0	0	0
0.15 to 0.24	0	0	0	0	0	0	0	0
0.25 to 0.34	0	0	0	0	0	0	0	0
0.35 to 0.44	0	0	0	0	0	0	0	0
0.45 to 0.54	0	0	0	0	0	0	0	0
≥ 0.55	100	4	100	1	100	4	100	4
Descriptive Stats								
Number of Items	4		1		4		4	
Mean	0.63		0.70		0.76		0.63	
SD	0.02		0.00		0.05		0.04	
Min	0.62		0.70		0.71		0.57	
Max	0.66		0.70		0.81		0.66	

Table 5.5 January Field Test Items Excluded from Analyses

	Algebra		Biology		Geometry		Government	
	SR	CR	SR*	CR	SR	CR	SR	CR
Not Scored	0	0	0	1	0	0	0	0
Low/Neg Pt-Biserial/Flawed	0	0	1	0	1	0	0	0
No Response for Highest Score Level	0	0	0	2	0	0	0	2

* One additional item was excluded. Item was operational, and changed to FT due to key change.

The item was calibrated, but with too few cases. Not banked for future use.