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, which is an ETS proprietary software. The analysis procedures for each component are described in detail. Samples used for the analyses included all valid records available, including students learning 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.

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 MDHSA field test items, and associated criteria used to flag items for the content specialists' review, 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. For SR items, the item-total correlation is the point-biserial correlation. For CR items, the item-total correlation is the polyserial correlation. 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, it is assumed 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. Summaries of p-values by content area for the field test items administered in May are found in Table 5.5 for SR items and Table 5.6 for CR items. Summaries of item-total correlations by content area for the field test items administered in May are found in Table 5.7 for the SR items and Table 5.8 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 2006 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.
- *Distracter Flag*: Point-biserial correlation positive for incorrect option.
- *Omit Flag*: Percent omitted is greater than 5.
- *Collapsed Score Levels*: Items with no students obtaining the score point.

Following the classical item analyses, most items with poor item statistics and all items that were not scored, as per MSDE's instructions, were removed from further analyses. Table 5.9 presents items that were removed from further analyses and identified for revision and possible re-field testing. Table 5.10 presents items that, although flagged for statistical reasons including high omit rates; extreme p-values; low correlations; missing responses for SR distracters or CR score points; and C-Level DIF, were retained for further analyses and evaluation. Calibration results indicated the items were estimated reasonably, and therefore were not removed from scaling.

Differential Item Functioning

Following the classical item analyses, differential item functioning (DIF) analyses were completed. One goal of test development is to assemble a set of items that provides an estimate of student 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. 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[4]. This statistic is expressed as the difference between the focal and reference group performance on an item after conditioning on total test score. 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 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 treated as the reference groups for gender and ethnicity, respectively; the female and other ethnic groups were considered the focal groups.

Based on their 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 (CR) items, the MH D-DIF statistic is not calculated; instead the standardization procedure is used in conjunction with the Mantel chi-square statistic. Analogous flagging rules have been developed that are used to classify the CR items into A, B, or C DIF categories. The flagging criteria for constructed response items are:

---

[4] 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):
$MH\ D\text{-}DIF = -2.35 \ln[\alpha_{MH}]$.

A) If the Mantel Chi-square p-value > 0.05 and/or the Mantel Chi-square p-value < 0.05 and the Standardized Mean Difference (SMD)/SD| ≤ 0.17, the item is classified as A.

B) If the Mantel Chi-square p-value < 0.05 and |SMD/SD| > 0.17 then the item is classified as B.

C) If the Mantel Chi-square p-value < 0.05 and |SMD/SD| > 0.25 then the item is classified as C.

Positive values favor the focal group and negative values favor the reference group.

There were 4 items flagged for C-level DIF against one of the identified focal groups (i.e., female, African American, American Indian, Asian, Hispanic). Refer to Table 5.10. The item in the January Algebra test was a CR item that favored females. The item in the May Algebra test was a SR item that was flagged for three focal groups of Hispanics, African Americans, and females, and favored the reference groups of Whites and males. The biology SR item in the January test favored males as opposed to females, as did the Government SR item in the May test. These items are flagged in the bank and will be reviewed.

IRT Calibration and Scaling

One 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.

As noted previously, the IRT models used to calibrate the MDHSA 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 discrimination ;
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 3PL model were discrimination (a), difficulty (b), and the pseudo-guessing level (c).

The GPCM is given by

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

where

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

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

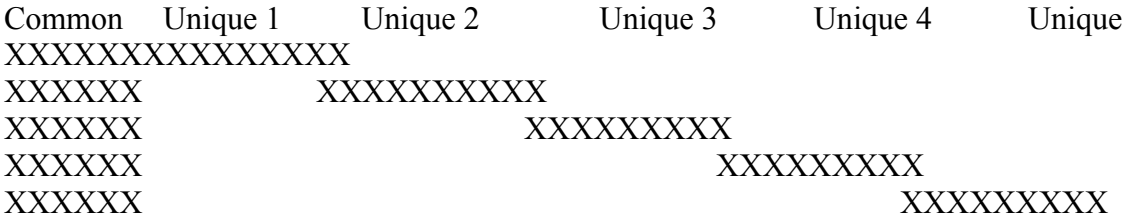
P_{jk} is the probability of responding in the k^{th} category from m_j+1 categories for item j,
θ is the ability level,
a_j is the item parameter characterizing the discrimination for item j,
b_{jk} is an item-category parameter for item j,
b_j is the item parameter characterizing the difficulty for item j, and
d_k is the category parameter characterizing the relative difficulty of categories 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 Stocking and Lord's (1983) TCC method and 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 set up 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.



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

	Percent and Number of Items					
P-Value	Algebra		Biology		Government	
	%	N	%	N	%	N
< 0.30[a]	6.25	1	12.50	3	0.00	0
0.30 to 0.40	12.50	2	20.83	5	20.00	2
0.41 to 0.50	18.75	3	8.33	2	40.00	4
0.51 to 0.60	12.50	2	25.00	6	10.00	1
0.61 to 0.70	12.50	2	4.17	1	10.00	1
0.71 to 0.80	18.75	3	25.00	6	20.00	2
≥ 0.81[b]	18.75	3	4.17	1	0.00	0
Descriptive Stats						
Number of Items	16		24		10	
Mean	0.59		0.53		0.52	
SD	0.21		0.19		0.17	
Min	0.18		0.16		0.30	
Max	0.94		0.84		0.78	

[a] p-value < 0.25: 1 algebra item; 1 biology item
[b] p-value > 0.90: 1 algebra item

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

	Percent and Number of Items					
P-Value	Algebra		Biology		Government	
	%	N	%	N	%	N
< 0.30	0.00	0	50.00	1	0.00	0
0.30 to 0.40	33.33	1	50.00	1	50.00	1
0.41 to 0.50	33.33	1	0.00	0	50.00	1
0.51 to 0.60	33.33	1	0.00	0	0.00	0
0.61 to 0.70	0.00	0	0.00	0	0.00	0
0.71 to 0.80	0.00	0	0.00	0	0.00	0
≥ 0.81	0.00	0	0.00	0	0.00	0
Descriptive Stats						
Number of Items	3		2		2	
Mean	0.46		0.30		0.38	
SD	0.05		0.03		0.03	
Min	0.40		0.27		0.36	
Max	0.51		0.32		0.41	

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

	Percent and Number of Items					
Correlation	Algebra		Biology		Government	
	%	N	%	N	%	N
< 0.15	0.00	0	12.50	3	0.00	0
0.15 to 0.24	0.00	0	16.67	4	30.00	3
0.25 to 0.34	31.25	5	16.67	4	40.00	4
0.35 to 0.44	50.00	8	20.83	5	30.00	3
0.45 to 0.54	12.5	2	33.33	8	0.00	0
≥ 0.55	6.25	1	0.00	0	0.00	0
Descriptive Stats						
Number of Items	16		24		10	
Mean	0.38		0.35		0.32	
SD	0.09		0.15		0.07	
Min	0.25		0.04		0.21	
Max	0.56		0.52		0.40	

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

	Percent and Number of Items					
Correlation	Algebra		Biology		Government	
	%	N	%	N	%	N
< 0.15	0.00	0	0.00	0	0.00	0
0.15 to 0.24	0.00	0	0.00	0	0.00	0
0.25 to 0.34	0.00	0	0.00	0	0.00	0
0.35 to 0.44	0.00	0	0.00	0	0.00	0
0.45 to 0.54	33.33	1	0.00	0	0.00	0
≥ 0.55	66.67	2	100.00	2	100.00	2
Descriptive Stats						
Number of Items	3		2		2	
Mean	0.60		0.66		0.71	
SD	0.08		0.01		0.02	
Min	0.52		0.65		0.69	
Max	0.68		0.67		0.72	

Table 5.5 Distribution of P-Values for the May Field Test SR Items

	Percent and Number of Items					
P-Value	Algebra		Biology		Government	
	%	N	%	N	%	N
< 0.30[a]	12.50	6	1.30	1	0.00	0
0.30 to 0.40	12.50	6	9.09	7	6.67	2
0.41 to 0.50	14.58	7	16.88	13	16.67	5
0.51 to 0.60	22.92	11	20.78	16	26.67	8
0.61 to 0.70	18.75	9	23.38	18	23.33	7
0.71 to 0.80	16.67	8	14.29	11	20.00	6
≥ 0.81[b]	2.08	1	14.29	11	6.67	2
Descriptive Stats						
Number of Items	48		77		30	
Mean	0.52		0.61		0.59	
SD	0.17		0.16		0.14	
Min	0.17		0.29		0.36	
Max	0.81		0.93		0.84	

[a] p-value < 0.25: 3 algebra items
[b] p-value > 0.90: 1 biology item

Table 5.6 Distribution of P-Values for the May Field Test CR Items

	Percent and Number of Items					
P-Value	Algebra		Biology		Government	
	%	N	%	N	%	N
< 0.30[a]	0.00	0	70.00	7	0.00	0
0.30 to 0.40	45.45	5	30.00	3	100.00	5
0.41 to 0.50	54.55	6	0.00	0	0.00	0
0.51 to 0.60	0.00	0	0.00	0	0.00	0
0.61 to 0.70	0.00	0	0.00	0	0.00	0
0.71 to 0.80	0.00	0	0.00	0	0.00	0
≥ 0.81	0.00	0	0.00	0	0.00	0
Descriptive Stats						
Number of Items	11		10		5	
Mean	0.41		0.28		0.36	
SD	0.06		0.05		0.05	
Min	0.32		0.18		0.31	
Max	0.50		0.33		0.40	

[a] p-value < 0.25: 1 biology item

Table 5.7 Distribution of Item-Total Correlations for the May Field Test SR Items

	Percent and Number of Items					
Correlation	Algebra		Biology		Government	
	%	N	%	N	%	N
< 0.15	4.17	2	0.00	0	3.33	1
0.15 to 0.24	10.42	5	5.19	4	13.33	4
0.25 to 0.34	29.17	14	24.68	19	16.67	5
0.35 to 0.44	29.17	14	45.45	35	26.67	8
0.45 to 0.54	25.00	12	24.68	19	36.67	11
≥ 0.55	2.08	1	0.00	0	3.33	1
Descriptive Stats						
Number of Items	48		77		30	
Mean	0.37		0.39		0.39	
SD	0.11		0.08		0.12	
Min	0.11		0.21		0.10	
Max	0.59		0.52		0.59	

Table 5.8 Distribution of Item-Total Correlations for May Field Test CR Items

	Percent and Number of Items					
Correlation	Algebra		Biology		Government	
	%	N	%	N	%	N
< 0.15	0.00	0	0.00	0	0.00	0
0.15 to 0.24	0.00	0	0.00	0	0.00	0
0.25 to 0.34	0.00	0	0.00	0	0.00	0
0.35 to 0.44	0.00	0	0.00	0	0.00	0
0.45 to 0.54	0.00	0	0.00	0	0.00	0
≥ 0.55	100.00	11	100.00	10	100.00	5
Descriptive Stats						
Number of Items	11		10		5	
Mean	0.67		0.70		0.71	
SD	0.04		0.04		0.02	
Min	0.61		0.64		0.69	
Max	0.73		0.77		0.74	

Table 5.9 Field Test Items Excluded from Analyses by Administration and Content Area

Content	MD ID	Form	Sequence	Response Type	Reason
January					
Algebra	79121	B	12	CR	grid incorrectly appeared in answer book;
					MSDE requested Do Not Score
Biology	79404	ACD	67	SR	R = -0.06
	79483	B	8	SR	R = -0.13
	79453	B	9	SR	R = -0.03
	79492	B	51	SR	R = -0.03
May					
Algebra	106647	F	30	CR	item content issue;
					MSDE requested Do Not Score
Biology	108504	K	8	SR	R = -0.04
Government	135493	H	48	CR	mistake in scoring (pooled with a similar item);
					MSDE requested Do Not Score

Table 5.10 Field Test Items with Statistical Flags Retained in Analysis

	Statistical Flag						
	Omit Rate		P-value		Corr	Missing responses for distracters or score points	C-Level DIF
	SR≥5%	CR≥20%	≤ 0.10	≥ 0.90	< 0.10		
January							
Algebra	2	0	0	1	0	0	1
Biology	0	0	0	0	2	0	1
Government	0	0	0	0	0	0	0
May							
Algebra	3	0	0	0	0	0	1
Biology	0	0	0	1	0	0	0
Government	0	0	0	0	0	0	1