Section 5. Field Test Analyses

Following the receipt of the final score 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 program. 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 records invalidated by the test administrator, and records with five or fewer item responses were excluded from the analysis sample.

The field test analyses were conducted for the SR and SPR items from the January and May administrations. The CR items were not scored or calibrated because beginning in May 2009, the operational forms will contain only SR and SPR items. Therefore field test CR item parameters were not needed for future administrations.

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 mean item score expressed as a proportion of the maximum obtainable item score. For SR and SPR items, it is equivalent to the proportion 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.

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

> This statistic describes the relationship between performance on the specific item and performance on the total test including the item under study. 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 raw score:

These statistics describe the relationship between selecting an incorrect response option for a specific item and performance on the total test including the item under study. 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. For individual items, if the omit percentage is greater than 5% for a single SR or SPR item or 15% for a CR item[8], this could be an indication of an

[8] Omit rates are typically greater for CR items than for SR/SPR items, therefore a higher omit percentage is used to signal a potential CR item/test layout problem.

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. Summaries of item-total correlations by content area for the field test items administered in January are found in Table 5.2. Summaries of p-values and item-total correlations by content area for the field test items administered in May are found in Tabled 5.3 and 5.4, respectively. 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 2008 assessments:

- *Difficulty Flag*: P-values less than 0.25 or greater than 0.90.
- *Discrimination Flag*: Item-total correlation less than 0.15.
- *Distractor Flag*: SR point-biserial correlation positive for incorrect option.
- *Omit Flag*: Percent omitted is greater than 5 for SR and SPR items and 15 for CR items.
- *Collapsed Score Levels*: Operational CR items with no students obtaining the score point.

Following the classical item analyses, items with poor item statistics and items that were not scored as per MSDE's instructions were removed from further analyses. Refer to Table 5.5. These items have been identified for revision and possible re-field testing. Table 5.6 presents the number of items that, although flagged for statistical reasons including extreme p-values; low item-total correlations; and/or high omits rates; 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. A subsequent review by MSDE and ETS content experts is conducted to investigate 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[9]. 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. 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

---

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

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

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):

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

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:

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| between 0.17 and 0.25 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.

None of the January field test items were flagged for DIF. For the May administration, forty-nine field test items were flagged for C-level DIF involving one or more of the identified focal groups (i.e., female, African American, American Indian, Asian, Hispanic). The numbers of items flagged for DIF were 7 Algebra items, 6 Biology items, 20 English items, and 16 Government items. These items are flagged in the item bank and will be reviewed by ETS and MSDE content specialists as well as ETS senior staff to determine their availability for future use.

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 3PL model relates the probability that a person with ability θ will respond correctly to item i as follows:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}}$$

where:

a_i is the slope parameter of item i, characterizing its discrimination ;

b_i is the location parameter of item i, characterizing its difficulty; and

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

The GPCM states that the probability that a person with ability θ will obtain a score of k on an item i_that has m score categories assigned score values ranging from 0 to m-1 can be expressed as

$$P_{ik}(\theta) = \frac{e^{Z_{ik}}}{1 + \sum_{k=1}^{m-1} e^{Z_{ik}}},$$

where

$$Z_{ik}(\theta) = 1.7a_i\theta - (\sum_{v=0}^{k} b_i - d_{iv}) = 1.7a_i\theta - \sum_{v=0}^{k} b_{iv}$$

$b_{i0} = 0$, and

P_{ik} is the probability of obtaining a score of k on item i, and

d_{ik} is the parameter characterizing the relative difficulty of obtaining score 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.

Common	Unique 1	Unique 2	Unique 3	Unique 4	Unique 5
XXXXXXXX	XXXXXXXX				
XXXXXXXX		XXXXXXXX			
XXXXXXXX			XXXXXXXX		
XXXXXXXX				XXXXXXXX	
XXXXXXXX					XXXXXXXX

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 banked parameters of the primary form operational non-CR 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 that were not calibrated were listed as unavailable (see Table 5.5).

Statistical Summary Tables

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

	Percentage and Number of Items							
P-Value	Algebra[a]		Biology		English		Government	
	%	N	%	N	%	N	%	N
P < 0.25	37.50	6	10.71	3	0.00	0	0.00	0
0.25 ≤ P < 0.35	12.50	2	17.86	5	9.38	3	20.00	2
0.35 ≤ P < 0.45	18.75	3	10.71	3	9.38	3	0.00	0
0.45 ≤ P < 0.55	18.75	3	35.71	10	18.75	6	50.00	5
0.55 ≤ P < 0.65	0.00	0	10.71	3	34.38	11	20.00	2
0.65 ≤ P < 0.75	12.50	2	7.14	2	21.88	7	10.00	1
0.75 ≤ P < 0.85	0.00	0	3.57	1	6.25	2	0.00	0
P ≥ 0.85	0.00	0	3.57	1	0.00	0	0.00	0
Descriptive Statistics								
N Items	16		25		32		10	
Mean	0.33		0.51		0.57		0.49	
SD	0.20		0.15		0.13		0.11	
Min	0.07		0.27		0.30		0.30	
Max	0.70		0.85		0.78		0.67	

[a] SPR items included

Table 5.2 Distribution of Item-Total Correlations for the January Field Test Items

	Percentage and Number of Items							
Correlation	Algebra[a]		Biology		English		Government	
	%	N	%	N	%	N	%	N
R < 0.15	12.50	2	17.86	5	0.00	0	0.00	0
0.15 ≤ R < 0.25	31.25	5	7.14	2	12.50	4	10.00	1
0.25 ≤ R < 0.35	43.75	7	10.71	3	25.00	8	50.00	5
0.35 ≤ R < 0.45	12.50	2	39.29	11	50.00	16	20.00	2
0.45 ≤ R < 0.55	0.00	0	25.00	7	12.50	4	20.00	2
0.55 ≤ R < 0.65	0.00	0	0.00	0	0.00	0	0.00	0
0.65 ≤ R < 0.75	0.00	0	0.00	0	0.00	0	0.00	0
R ≥ 0.75	0.00	0	0.00	0	0.00	0	0.00	0
Descriptive Statistics								
N Items	16		25		32		10	
Mean	0.25		0.37		0.36		0.35	
SD	0.10		0.13		0.08		0.10	
Min	0.01		0.02		0.18		0.19	
Max	0.41		0.53		0.49		0.52	

[a] SPR items included

Table 5.3 Distribution of P-Values for the May Field Test Items

	Percentage and Number of Items							
P-Value	Algebra[a]		Biology		English		Government	
	%	N	%	N	%	N	%	N
P < 0.25	7.33	11	1.44	4	2.12	6	1.67	5
0.25 ≤ P < 0.35	12.67	19	5.42	15	4.95	14	5.67	17
0.35 ≤ P < 0.45	13.33	20	13.36	37	7.07	20	14.33	43
0.45 ≤ P < 0.55	22.67	34	22.38	62	13.43	38	14.00	42
0.55 ≤ P < 0.65	12.00	18	21.30	59	20.49	58	24.33	73
0.65 ≤ P < 0.75	18.67	28	19.13	53	22.97	65	19.67	59
0.75 ≤ P < 0.85	8.67	13	13.72	38	24.73	70	12.00	36
P ≥ 0.85[b]	4.67	7	3.25	9	4.24	12	8.33	25
Descriptive Statistics								
N Items	150		277		283		300	
Mean	0.52		0.58		0.63		0.60	
SD	0.19		0.16		0.16		0.17	
Min	0.13		0.18		0.12		0.13	
Max	0.92		0.92		0.95		0.92	

[a] SPR items included; [b] P-value > 0.90: 2 Algebra, 1 Biology , 3 English, and 2 Government items

Table 5.4 Distribution of Item-Total Correlations for the May Field Test Items

	Percentage and Number of Items							
Correlation	Algebra[a]		Biology		English		Government	
	%	N	%	N	%	N	%	N
R < 0.15	1.33	2	4.33	12	6.71	19	4.33	13
0.15 ≤ R < 0.25	5.33	8	11.19	31	14.49	41	6.67	20
0.25 ≤ R < 0.35	19.33	29	21.30	59	30.04	85	20.33	61
0.35 ≤ R < 0.45	37.33	56	40.79	113	36.40	103	37.00	11
0.45 ≤ R < 0.55	25.33	38	21.30	59	12.01	34	29.33	88
0.55 ≤ R < 0.65	8.00	12	1.08	3	0.35	1	2.33	7
0.65 ≤ R < 0.75	3.33	5	0.00	0	0.00	0	0.00	0
R ≥ 0.75	0.00	0	0.00	0	0.00	0	0.00	0
Descriptive Statistics								
N Items	150		277		283		300	
Mean	0.42		0.36		0.33		0.38	
SD	0.12		0.11		0.11		0.12	
Min	0.01		-0.01		-0.16		-0.11	
Max	0.69		0.61		0.55		0.59	

[a] SPR items included

Table 5.5 Field Test Items Excluded from Calibration

Administration	Content	ItemID	Form	Sequence	Response Type	Reason
January						
	Algebra	106600	A	17	CR	MSDE requested Do Not Score
	Algebra	136696	A	29	SR	R_ITT=0.01
	Algebra	79530	A	30	CR	MSDE requested Do Not Score
	Algebra	79125	B	17	CR	MSDE requested Do Not Score
	Algebra	106538	B	30	CR	MSDE requested Do Not Score
	Biology	108678	A	11	CR	MSDE requested Do Not Score
	Biology	137272	A	13	SR	R_ITT=0.02
	Biology	108667	A	62	CR	MSDE requested Do Not Score
	Biology	215974	B	5	SR	MSDE requested Do Not Score
	Biology	215975	B	6	SR	MSDE requested Do Not Score
	Biology	135582	B	11	CR	MSDE requested Do Not Score
	Biology	215982	B	13	SR	MSDE requested Do Not Score
	Biology	223399	B	62	CR	MSDE requested Do Not Score
	English	251085	A	33	CR	MSDE requested Do Not Score
	English	251253	B	34	CR	MSDE requested Do Not Score
	Government	79557	A	29	CR	MSDE requested Do Not Score
	Government	79536	B	29	CR	MSDE requested Do Not Score
May						
	Algebra	135104	K	29	SR	R_ITT=0.06
	Algebra	211052	M	40	SR	R_ITT=0.01
	Biology	256540	E	4	SR	R_ITT=0.05
	Biology	256526	F	51	SR	R_ITT=-0.01
	Biology	133022	L	4	SR	R_ITT=0.01
	Biology	241125	M	51	SR	R_ITT=0.07
	English	246668	F	9	SR	R_ITT=0.05
	English	246671	F	12	SR	R_ITT=0.04
	English	215779	F	27	SR	R_ITT=0.07
	English	215794	G	9	SR	R_ITT=-0.01
	English	215798	G	10	SR	R_ITT=0.04
	English	214698	H	54	SR	R_ITT=0.02
	English	256306	K	23	SR	R_ITT=0.04
	English	215806	K	69	SR	R_ITT=0.04
	English	215770	L	27	SR	R_ITT=0.07
	English	256293	L	70	SR	R_ITT=-0.16
	English	223338	M	11	SR	R_ITT=0.06
	English	223331	M	31	SR	R_ITT=0.02

Administration	Content	ItemID	Form	Sequence	Response Type	Reason
	English	218610	N	30	SR	R_ITT=0.04
	Government	52194	E	14	SR	R_ITT=-0.08
	Government	79706	E	65	SR	R_ITT=-0.11
	Government	214579	M	39	SR	R_ITT=0.03
	Government	214503	N	4	SR	R_ITT=0.07

Table 5.6 Field Test Items with Statistical Flags Retained in Calibration

	P-Value	P-Value	R_ITT	Distractor Pt-Bis	Omit Rate	C-Level DIF	Missing Response[a]	Total Flags	N Items[b]
	< 0.25	> 0.90	< 0.15	> 0	> 5%				
January									
Algebra	6	0	1	3	4	0	0	14	7
Biology	0	0	1	4	0	0	0	5	4
English	0	0	0	1	0	0	0	1	1
Government	0	0	0	2	0	0	0	2	2
May									
Algebra	9	2	0	5	28[c]	7	0	51	43
Biology	3	1	8	28	0	6	0	46	36
English	1	3	6	28	0	20	0	58	50
Government	4	2	9	25	0	16	0	56	45

[a] SR option with 0 students; [b] Represents total number of unique items; [c] All SPR items.