Section 5. Field Test Analyses

Following the receipt of the final scored file from Measurement Incorporated (MI), the field test analyses were completed. The analysis of the field test data can be broken down in four componentsThere are four types of analyses conducted for the field test data: 1) classical item analyses; 2) differential item functioning (DIF) analyses; and 3) calibration and scaling. All of the analyses were completed using Genasys, ETS proprietary software. The analysis procedures for each component are described in detail. The samples used for all analyses included all valid records available at the time of the analyses. Students classified as English as a second language, students with IEP or 504 plans and those receiving accommodations were included in all analyses. Only duplicate records, records invalidated by the test administrator and those with five or fewer item responses were excluded from the analysis sample.

Classical Item Analyses

Classical item analyses involve computing, for every item in each form, a set of statistics based on classical test theory. Each statistic is designed to provide some key information about the quality of each item from an empirical perspective. The statistics estimated for the HSA 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 to better measure students with very high or low achievement, especially if the students have not yet received instruction in the content or if they lack motivation to complete the field test items to the best of their ability.

The item-total correlation of the correct response option (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., incorrect key, multiple correct answers or unusually complex content), or can indicate that students have not been taught the content.

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

These statistics indicate the percent of examinees that select each of the available answer options and the percent of examinees that omitted the

item. Item options not selected by any students indicate problems with plausibility of the option. Items that do not have all answer options functioning should 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 of this correlation 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 can also 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 that 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 in identifying how well the item is functioning. If no students are assigned the top score point, this indicates that the item may not be functioning with respect to the rubric and/or that the is with no students can indicate serious problems with the item content or can indicate that students have not been taught the content.

Summaries of the items administered based on p-values are listed in and item-total correlations are listed in Tables 5.1-5.8 for each content area. In addition, a series of flags were created in order 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 was applied to all items tested in the 2003-2004 assessments:

- Difficulty Flag: P-values less than 0.25 or greater than 0.90.
- Discrimination Flag: Point- biserial correlation less than 0.15 for correct answer.

- Distractor Flag: Point-biserial correlation is 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 (see Tables 5.9 and 5.10). These items have been identified for revision and possible future re-field testing.

Differential Item Functioning (DIF)

Following the classical item analyses, DIF studies were completed. One of the goals 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 those items that identifiable groups of students (e.g. females, African Americans, Hispanics) with the same underlying level of ability have different probabilities of answering correctly. If the item is differentially more difficult for an identifiable subgroup, the item may be measuring something different from 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 are required to determine the source and meaning of any differences that are seen.

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

[9] 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):

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

differences between the focal and reference group after conditioning on total test score. This 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. 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 as reference groups and the female and other ethnic groups are categorized as 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 contains negligible DIF, Category B items exhibit slight or moderate DIF, and Category C items have moderate to large values of 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.

No items were flagged for C-level DIF against one of the identified focal groups (female, African American, American Indian, Asian, and Hispanic) for both January and May administrations.

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 versions within 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 HSA test items were the 3-parameter logistic (3PL) model for selected response items and the generalized partial credit model (GPCM) for constructed response 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

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

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 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
XXXXXXXXXXXXXXX
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. In some cases, there may be several iterations of calibrations whereby items that do not converge are removed from analysis. No items were omitted from the final calibrations.

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
- Collapsed score level
- Item not scored

Government Constructed Response Study

In the evolution of the item writing process, the directional statements associated with the Government brief and extended constructed response items were modified to be more specific, beginning with the May, 2004 administration. In reviewing the item bank, there were several items that could be used on future forms, however, these items included the previous directional statements and formatting. As a result, available items have two different formats and future test forms could include items with both types of formatting. While changing all of the items to the "new" format would be desirable, MSDE was concerned that this change could impact item performance. To obtain new item parameters, the items would need to be re-field tested, which would decrease the numbers of items available for form construction in the short term, would delay the field testing of newly written items, and would increase the development costs associated with these existing items. A study completed during the May 2004 administration that involved printing two items in both the old and new formats found that there were virtually no differences between the two sets of item parameters. Therefore the change in the directions does not appear to have an important or systematic effect on item performance (see Appendix 5.A).

Statistical Summary Tables

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

	Percentage of items (N)				
P-Values	Algebra	Biology	English I	Geometry	Government
< 0.30	62.5 (10)	17.9 (5)	2.9 (1)	75.0 (12)	16.7 (4)
0.30 to 0.40	25.0 (4)	21.4 (6)	20.6 (7)	18.8 (3)	20.8 (5)
0.41 to 0.50	6.3 (1)	28.6 (8)	2.9 (1)	6.3 (1)	12.5 (3)
0.51 to 0.60	0	10.7 (3)	26.5 (9)	0	20.8 (5)
0.61 to 0.70	0	7.1 (2)	23.5 (8)	0	20.8 (5)
0.71 to 0.80	6.3 (1)	7.1 (2)	20.6 (7)	0	8.3 (2)
> 0.81	0	7.1 (2)	2.9 (1)	0	0
Number of Items	16	28	34	16	24
Mean	0.29	0.47	0.57	0.26	0.48
SD	0.16	0.19	0.16	0.09	0.15
Min	0.11	0.19	0.24	0.12	0.20
Max	0.76	0.85	0.83	0.47	0.73

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

	Percentage of items (N)				
P-Values	Algebra	Biology	English I	Geometry	Government
< 0.30	75.0 (3)	100.0 (3)	0	100.0 (3)	83.3 (5)
0.30 to 0.40	25.0 (1)	0	50.0 (1)	0	16.7 (1)
0.41 to 0.50	0	0	50.0 (1)	0	0
0.51 to 0.60	0	0	0	0	0
0.61 to 0.70	0	0	0	0	0
0.71 to 0.80	0	0	0	0	0
> 0.81	0	0	0	0	0
Number of Items	4	3	2	3	6
Mean	0.23	0.14	0.41	0.24	0.22
SD	0.05	0.04	0.03	0.04	0.07
Min	0.18	0.11	0.39	0.19	0.14
Max	0.30	0.18	0.43	0.26	0.32

* Table information does not include items with collapsed levels

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

	Percentage of items (N)				
Correlation	Algebra	Biology	English I	Geometry	Government
< 0.15	37.5 (6)	7.1 (2)	2.9 (1)	31.3 (5)	8.3 (2)
0.15 to 0.24	25.0 (4)	10.7 (3)	11.8 (4)	31.3 (5)	29.2 (7)
0.25 to 0.34	25.0 (4)	42.9 (12)	14.7 (5)	6.3 (1)	8.3 (2)
0.35 to 0.44	12.5 (2)	28.6 (8)	17.6 (6)	18.8 (3)	12.5 (3)
0.45 to 0.54	0	10.7 (3)	52.9 (18)	6.3 (1)	33.3 (8)
> 0.55	0	0	0	6.3 (1)	8.3 (2)
Number of SR Items	16	28	34	16	24
Mean	0.20	0.33	0.40	0.24	0.35
SD	0.13	0.11	0.12	0.18	0.16
Min	-0.02	0.003	0.06	-0.10	-0.08
Max	0.45	0.52	0.54	0.56	0.57

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

	Percentage of items (N)				
Correlation	Algebra	Biology	English I	Geometry	Government
< 0.15	0	0	0	0	0
0.15 to 0.24	0	0	0	0	0
0.25 to 0.34	0	0	0	0	0
0.35 to 0.44	0	0	0	0	0
0.45 to 0.54	0	0	0	0	0
> 0.55	100.0 (4)	100.0 (3)	100.0 (2)	100.0 (3)	100.0 (6)
Number of Items	4	3	2	3	6
Mean	0.69	0.75	0.72	0.71	0.72
SD	0.05	0.02	0.01	0.12	0.03
Min	0.63	0.73	0.72	0.57	0.69
Max	0.74	0.76	0.73	0.81	0.77

* Table information does not include items with collapsed levels

Table 5.5. Distributions of P-Values for May Field Test SR Items

	Percentage of items (N)				
P-Values	Algebra	Biology	English I	Geometry	Government
< 0.30	15.6 (10)	8.0 (9)	0.6 (1)	28.1 (18)	5.3 (4)
0.30 to 0.40	15.6 (10)	17.9 (20)	10.6 (17)	17.2 (11)	16.0 (12)
0.41 to 0.50	17.2 (11)	22.3 (25)	16.3 (26)	14.1 (9)	21.3 (16)
0.51 to 0.60	25.0 (16)	22.3 (25)	26.3 (42)	18.8 (12)	22.7 (17)
0.61 to 0.70	14.1 (9)	12.5 (14)	20.6 (33)	9.4 (6)	21.3 (16)
0.71 to 0.80	9.4 (6)	16.1 (18)	22.5 (36)	10.9 (7)	13.3 (10)
> 0.81	3.1 (2)	0.9 (1)	3.1 (5)	1.6 (1)	0
Number of Items	64	112	160	64	75
Mean	0.50	0.51	0.58	0.43	0.53
SD	0.18	0.15	0.14	0.20	0.15
Min	0.11	0.16	0.20	0.05	0.16
Max	0.88	0.84	0.86	0.87	0.79

Table 5.6. Distributions of P-Values for May Field Test CR Items

	Percentage of items (N)				
P-Values	Algebra	Biology	English I**	Geometry	Government
< 0.30	53.3 (8)	56.3 (9)		46.7 (7)	76.9 (10)
0.30 to 0.40	6.7 (1)	12.5 (2)		13.3 (2)	23.1 (3)
0.41 to 0.50	20.0 (3)	0		33.3 (5)	0
0.51 to 0.60	0	0		6.7 (1)	0
0.61 to 0.70	0	0		0	0
0.71 to 0.80	0	0		0	0
> 0.81	0	0		0	0
Number of Items	12	11		15	13
Mean	0.28	0.24		0.35	0.27
SD	0.10	0.07		0.09	0.06
Min	0.11	0.14		0.24	0.15
Max	0.46	0.38		0.53	0.37

* Table information does not include items with collapsed levels

** No CR items were scored

Table 5.7. Distributions of Item-Total Correlations for May Field Test SR Items

	Percentage of items (N)				
Correlation	Algebra	Biology	English I	Geometry	Government
< 0.15	6.0 (4)	5.4 (6)	2.5 (4)	12.5 (8)	5.3 (4)
0.15 to 0.24	14.1 (9)	13.4 (15)	4.4 (7)	9.4 (6)	14.7 (11)
0.25 to 0.34	26.6 (17)	28.6 (32)	20.6 (33)	9.4 (6)	13.3 (10)
0.35 to 0.44	31.3 (20)	40.2 (45)	37.5 (60)	35.9 (23)	30.7 (23)
0.45 to 0.54	12.5 (8)	12.5 (14)	33.1 (53)	14.1 (9)	30.7 (23)
> 0.55	9.4 (6)	0.0 (0)	1.9 (3)	18.8 (12)	5.3 (4)
Number of SR Items	64	112	160	64	75
Mean	0.36	0.34	0.40	0.39	0.38
SD	0.13	0.10	0.10	0.16	0.13
Min	0.09	0.06	-0.04	0.01	0.02
Max	0.65	0.53	0.57	0.72	0.57

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

	Percentage of items (N)				
Correlation	Algebra	Biology	English I**	Geometry	Government
< 0.15	0	0		0	0
0.15 to 0.24	0	0		0	0
0.25 to 0.34	0	0		0	0
0.35 to 0.44	0	0		0	0
0.45 to 0.54	8.3 (1)	0		0	0
> 0.55	91.7 (11)	100.0 (11)		100.0 (15)	100.0 (13)
Number of CR Items	12	11		15	13
Mean	0.65	0.68		0.71	0.71
SD	0.06	0.04		0.07	0.05
Min	0.53	0.63		0.55	0.62
Max	0.74	0.75		0.79	0.78

* Table information does not include items with collapsed levels

** No CR items were scored

Table 5.9 Field Test Items Excluded from Analyses: January

	Algebra		Biology		English I		Geometry		Government	
	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR
Not Scored				1		1		1		
Low/Neg. P-biserial	1		1		1		1			
Collapsed levels		1		3						2

Table 5.10 Field Test Items Excluded from Analyses: May

	Algebra		Biology		English I		Geometry		Government	
	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR
Not Scored		3				9*		1		
Low/Neg. P-biserial			1		1				1	
Collapsed levels		1		6		1				5

* English 10 test will be replacing English I test after May 2005 administration so no English I field test CR items were scored for May 2004 administration.