Section 7. Field Test Analyses

Following the receipt of the final score file from Pearson for the January and May administrations, the field test analyses for SR and SPR items were completed. The analyses consisted of four components: classical item analyses, differential item functioning (DIF), item response theory (IRT) calibration, and scaling. All the analyses were completed using *GENASYS*, an ETS proprietary software program. The analysis procedures for each component are described in detail below. All valid records available were used as samples for the analyses, including those for 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 no item responses to the first five items 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 HSA 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.
>
> Item-total correlation of the correct response option for SR and SPR items: 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 and SPR items, the item-total correlation is the point-biserial correlation. Values less than 0.20 are generally considered to have a weaker than desired relationship, therefore these items receive careful review by ETS staff and MSDE before including them on future forms. Items with negative correlations can indicate there are serious problems with the item content (e.g., multiple correct answers, unusually complex content), there is an incorrect key, or students have not been taught the content.
>
> Proportion of students choosing each response option (SR items): This statistic indicates the percent of examinees selecting each answer choice, or option. 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.

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 options 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, at least 95 percent of them should attempt to answer each question. When a pattern of omit percentages exceeds 5 percent 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 percent for a single SR or SPR 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.

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 2012 assessments:

- *Difficulty flag*: *p*-values less than 0.10 or greater than 0.90.
- *Discrimination flag*: Item-total correlation less than 0.10.
- *Distractor flag*: SR point-biserial correlation positive for incorrect option.
- *Omit flag*: Percent omitted is greater than 5 for SR and SPR items.

Distributions of *p*-values and item-total correlations for the field test items administered in January 2012 are shown in Tables 7.1 and 7.2, respectively. Corresponding results for the field test items administered in May 2012 are shown in Tables 7.3 and 7.4, respectively.

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 (see Table 7.5). These items have been identified for revision and possible additional field testing. Table 7.6 presents the number of items that were retained for further analyses and evaluation after being flagged for statistical reasons, including extreme *p*-values, low item-total correlations, and/or high omit rates. Calibration results indicated the items were estimated reasonably; therefore they were not removed from scaling.

Table 7.1 Distribution of *p* -Values for the MD HSA January 2012 Field Test Items

	Percentage and Number of Items					
	Algebra[a]		Biology		English	
p-Value	%	N	%	N	%	N
$p < 0.25$	38	12	2	1	9	7
$0.25 \leq p < 0.35$	19	6	7	3	16	12
$0.35 \leq p < 0.45$	16	5	24	11	22	17
$0.45 \leq p < 0.55$	16	5	26	12	21	16
$0.55 \leq p < 0.65$	3	1	28	13	20	15
$0.65 \leq p < 0.75$	6	2	11	5	8	6
$0.75 \leq p < 0.85$	3	1	2	1	3	2
$p \geq 0.85$	0	0	0	0	1	1
Descriptive Statistics						
N Items	32		46		76	
Mean	0.32		0.50		0.46	
SD	0.20		0.13		0.16	
Min	0.04		0.23		0.11	
Max	0.77		0.78		0.86	

[a] SPR items included

Table 7.2 Distribution of Item-Total Correlations for the MD HSA January 2012 Field Test Items

	Percentage and Number of Items					
	Algebra[a]		Biology		English	
Correlation	%	N	%	N	%	N
$r < 0.15$ [c]	22	7	4	2	18	14
$0.15 \leq r< 0.25$	16	5	2	1	32	24
$0.25 \leq r < 0.35$	47	15	20	9	21	16
$0.35 \leq r < 0.45$	16	5	41	19	28	21
$0.45 \leq r < 0.55$	0	0	33	15	1	1
$0.55 \leq r < 0.65$	0	0	0	0	0	0
$0.65 \leq r < 0.75$	0	0	0	0	0	0
$r \geq 0.75$	0	0	0	0	0	0
Descriptive Statistics						
N Items	32		46		76	
Mean	0.25		0.40		0.26	
SD	0.11		0.10		0.13	
Min	-0.04		0.13		-0.09	
Max	0.44		0.54		0.47	

[a] SPR items included; [c] $r < 0.10$: 4 Algebra, and 7 English items

Table 7.3 Distribution of p -Values for the MD HSA May 2012 Field Test Items

	Percentage and Number of Items					
	Algebra[a]		Biology		English	
p -Value	%	N	%	N	%	N
$p < 0.25$	11	18	0	1	6	21
$0.25 \leq p < 0.35$	9	15	1	3	7	28
$0.35 \leq p < 0.45$	17	27	11	25	16	61
$0.45 \leq p < 0.55$	20	32	25	57	18	70
$0.55 \leq p < 0.65$	15	24	26	60	19	74
$0.65 \leq p < 0.75$	16	26	19	44	16	59
$0.75 \leq p < 0.85$	8	12	12	28	15	56
$p \geq 0.85$[b]	4	6	5	12	3	11
Descriptive Statistics						
N Items	160		230		380	
Mean	0.51		0.60		0.55	
SD	0.19		0.14		0.18	
Min	0.09		0.23		0.12	
Max	0.92		0.97		0.89	

[a] SPR items included; [b] p -value > 0.90: 1 Algebra, and 3 Biology

Table 7.4 Distribution of Item-Total Correlations for the MD HSA May 2012 Field Test Items

	Percentage and Number of Items					
	Algebra[a]		Biology		English	
Correlation	%	N	%	N	%	N
$r < 0.15$[c]	3	4	1	3	13	51
$0.15 \leq r < 0.25$	7	11	5	12	19	72
$0.25 \leq r < 0.35$	20	32	19	43	26	97
$0.35 \leq r < 0.45$	41	66	44	101	28	107
$0.45 \leq r < 0.55$	22	35	28	65	14	52
$0.55 \leq r < 0.65$	8	12	3	6	0	1
$0.65 \leq r < 0.75$	0	0	0	0	0	0
$r \geq 0.75$	0	0	0	0	0	0
Descriptive Statistics						
N Items	160		230		380	
Mean	0.39		0.40		0.30	
SD	0.11		0.10		0.14	
Min	0.03		0.02		-0.15	
Max	0.62		0.59		0.55	

[a] SPR items included; [c] $r < 0.10$: 4 Algebra, 1 Biology, and 34 English items

Table 7.5 MD HSA Field Test Items Excluded from Calibration

Administration	Content	ItemID	Form	Sequence	Response Type[a]	Reason[b]
January	Algebra	369144	A	48	SR	Rbis = 0.09
	Algebra	408223	A	51	SR	Rbis = -0.05
	Algebra	257134	B	63	SR	Rbis = 0.05
	English	364048	A	96	SR	Rbis = 0.07
	English	369280	B	7	SR	Rbis = -0.15
	English	369279	B	8	SR	Rbis = -0.07
	English	369278	B	9	SR	Rbis = 0.00
	English	369282	B	10	SR	Rbis = -0.03
May	Algebra	393647	E	3	SR	Rbis= 0.04
	Algebra	393644	G	18	SPR	Do Not Score : invalid content
	Algebra	393688	G	44	SR	Rbis= 0.08
	Biology	394746	J	78	SR	Rbis= 0.02
	English	397713	D	11	SR	Rbis= -0.02
	English	392559	D	26	SR	Rbis= -0.04
	English	364053	D	49	SR	Rbis= 0.04
	English	369019	D	73	SR	Rbis= -0.04
	English	363069	D	80	SR	Rbis= -0.10
	English	373705	E	21	SR	Rbis= 0.09
	English	373708	E	24	SR	Rbis= -0.10
	English	366280	E	79	SR	Rbis= -0.18
	English	363111	E	80	SR	Rbis= 0.04
	English	397731	F	9	SR	Rbis= 0.09
	English	363082	F	83	SR	Rbis= 0.00
	English	369294	G	71	SR	Rbis= 0.04
	English	369299	G	75	SR	Rbis= -0.22
	English	359859	G	81	SR	Rbis= -0.09
	English	364110	K	77	SR	Rbis= 0.08
	English	366311	L	71	SR	Rbis= 0.04
	English	366321	L	79	SR	Rbis= 0.05
	English	393698	M	21	SR	Rbis= 0.07
	English	364009	M	75	SR	Rbis= -0.03
	English	364035	M	83	SR	Rbis= 0.01
	English	364065	N	73	SR	Rbis= -0.16
	English	364059	N	77	SR	Rbis= 0.09

[a] SR = Selected-response item; SPR = Student-produced response item; [b] Rbis = Biserial correlation

Table 7.6 MD HSA Field Test Items with Statistical Flags Retained in Calibration

January	*p*-Value	*p*-Value	Point Biserial	Distractor Pt-Bis	Omit Rate	C-Level DIF	Missing Response[a]	Total Flags	N Items[b]
	< 0.10	> 0.90	< 0.10	> 0	> 5%				
Algebra	6	0	1	5	5	3	0	20	14
Biology	0	0	0	2	0	0	0	2	2
English	0	0	2	20	0	4	0	26	24
May	*p*-Value	*p*-Value	Point Biserial	Distractor Pt-Bis	Omit Rate	C-Level DIF	Missing Response[a]	Total Flags	N Items[b]
	< 0.10	> 0.90	< 0.10	> 0	> 5%				
Algebra	0	1	2	9	9	2	0	23	20
Biology	0	3	0	6	0	5	0	14	14
English	0	0	12	76	0	15	0	103	91

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

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 (focal) groups of students with the same underlying level of ability (e.g., Females, African Americans, Asians, and Hispanics) have different probabilities of answering correctly than reference groups (Males, White students). If the item is 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. A subsequent review by MSDE and ETS content experts is conducted to investigate the source and meaning of evident differences.

ETS used the Mantel-Haenszel DIF detection method. As part of the Mantel-Haenszel procedure, the statistic described by Holland & Thayer (1988), known as MH D-DIF, was used[6]. This statistic is expressed as the difference between the focal and reference group

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

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

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 and 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 race and 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 a lower mean item score than the focal group.

Among the items field-tested in January, three items for Algebra, none in Biology, and four English items were flagged for C-level DIF. Among the items field tested in May, two items for Algebra, five Biology items, and fifteen English items were flagged for C-level DIF. These flags were recorded in the item bank. Flagged items are reviewed by ETS and MSDE content specialists as well as by ETS senior staff to determine their suitability 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. 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

---

where
RB_{rmB} = number in reference group at ability level *m* answering the item right,
WB_{fmB} = number in focal group at ability level *m* answering the item wrong,
RB_{fmB} = number in focal group at ability level *m* answering the item right,
WB_{rmB} = number in reference group at ability level *m* answering the item wrong,
NB_{mB} = total group at ability level *m*.

This can then be used in the following formula (Holland & Thayer, 1988):

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

known as a scale score), which can be more meaningfully interpreted by students, teachers, and other stakeholders.

As noted previously, the IRT model used to calibrate the MD HSA test items is the 3-parameter logistic (3PL) model. 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.

A proprietary version of the *PARSCALE* computer program (Muraki & Bock, 1995) was used for all item calibration work. The resulting calibrations were then scaled to the bank estimates using Stocking and Lord’s (1983) test characteristic curve (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, all items were calibrated using a sparse matrix design that places all items on a common scale. Essentially, this means that the data were set up using the following format. In the diagram below, X’s represent items and spaces indicate missing data. For example, items included on version 2 but not on version 1, 3, 4, or 5 were treated as “not administered” .

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 were reviewed to determine if any items failed to correctly calibrate.

3. After the final calibration, parameter estimates were obtained. The items were then linked to the bank scale using the TCC method. Specifically, the banked parameter estimates of the primary form 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, they were loaded into the item bank. Items that were not calibrated were listed as unavailable (see Table 7.5).