Section 3. Item-Level Analyses: May Administration

Analyses of the field test items were conducted following receipt of the final scored student data files. Item analyses results were examined prior to the selection of operational items. Item-level analyses consisted of classical item analyses and differential item functioning (DIF). Analyses were completed using GENASYS.

As mentioned in the introduction, two groups of students were administered the Mod-HSAs during the May administration. Item analyses and DIF were conducted separately for the *Target* population, students identified as being eligible to take the Mod-HSAs, and for the *Linking* samples that took both the Mod-HSA and the HSA.

Data Files

The data 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.

For the Target population who could take the Mod-HSAs online or in paper format, data were combined across mode of administration for each form, for all analyses.

Results of the item analyses and DIF analyses were provided to MSDE in Excel files containing item-level statistics, by form, for each content area. The files included blueprint information, classical item statistics, DIF statistics and flags for item statistics outside of the range of criteria approved by MSDE's technical advisors, National Psychometrics Council (NPC). These criteria are described later in this section. Also included in the files was a flag which indicated whether an item had been originally selected as part of the 50-item base form described in Section 1 of this report.

While data provided by the Linking samples were used to select operational items, statistics based on the Target population were also included in the files so that results from the two groups of students could be compared. To assist MSDE in their selection of operational forms, items flagged with statistics outside the range of the criteria for the Linking sample students were highlighted in red. Items with acceptable statistics for the Linking samples but less than desirable statistics for the Target population were highlighted in green and flagged as "Use with Caution." A variable indicating the number of items required for each subscore was also included in the Excel files.

Classical Item Analyses

Classical item analyses involve computing a set of statistics 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 Mod-HSA items, and associated criteria used to flag items for 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 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 0.10 to 0.90. Occasionally, items that fall outside this range can be justified for inclusion as an operational item 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:

> 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. Values less than 0.10 were flagged for a weaker than desired relationship and requiring careful consideration by MSDE before including them on operational 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:

> 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 did not have all answer options functioning would require careful consideration by MSDE before including on operational forms.

The point-biserial correlation of incorrect response option with the total 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, this could be an indication of an item/test layout problem. For example, students might accidentally skip an item that follows a lengthy stem.

The P-values for all of the Mod-HSA items administered are summarized for the Linking samples and for the Target populations in Tables 3.1 and 3.2. The point-biserials for these items and groups are summarized in Table 3.3 and 3.4. Recall that statistics from the Linking samples were used to select the operational 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 to place the Mod-HSA operational items onto the HSA scale. The following flagging criteria were applied to all Mod-HSA items administered in May 2008:

- *Difficulty Flag*: P-value less than 0.10 or greater than 0.90.
- *Discrimination Flag*: Point-biserial correlation less than 0.10 for the correct answer.
- *Distractor Flag*: Positive point-biserial correlation for incorrect option.
- *Omit Flag*: Percent omitted is greater than 5.

Following classical item analyses, items with poor item statistics were removed from further analyses (refer to Table 3.5). While these items were retained in the Mod-HSA item bank, they have been identified as "Do Not Use." Table 3.6 presents the number of items that were flagged but retained for further analyses and evaluation. These items were 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, and therefore were not removed from scaling.

Table 3.1 Distributions of P-Values: May All Mod-HSA Items – Linking

	Number and Percentage of Items							
P-Value	Algebra		Biology		English		Government	
	N	%	N	%	N	%	N	%
$P < 0.10$	0	0.00	2	1.46	0	0.00	0	0.00
$0.10 \leq P < 0.20$	0	0.00	1	0.73	0	0.00	0	0.00
$0.20 \leq P < 0.30$	2	1.33	1	0.73	0	0.00	0	0.00
$0.30 \leq P < 0.40$	6	4.00	3	2.19	4	3.48	0	0.00
$0.40 \leq P < 0.50$	9	6.00	7	5.11	3	2.61	4	2.67
$0.50 \leq P < 0.60$	16	10.67	16	11.68	12	10.43	15	10.00
$0.60 \leq P < 0.70$	24	16.00	29	21.17	17	14.78	26	17.33
$0.70 \leq P < 0.80$	41	27.33	36	26.28	31	26.96	44	29.33
$0.80 \leq P < 0.90$	33	22.00	30	21.90	35	30.43	37	24.67
$P \geq 0.90$	19	12.67	14	10.22	13	11.30	24	16.00
Descriptive Statistics								
N Items*	150		137		115		150	
Mean	0.72		0.71		0.74		0.76	
SD	0.16		0.16		0.14		0.13	
Min	0.28		0.13		0.30		0.42	
Max	0.98		0.96		0.97		0.99	

* N Items includes the number of unique items; some Biology and English items appear on both Forms 108 and 208.

Table 3.2 Distributions of P-Values: May All Mod-HSA Items – Target

	Number and Percentage of Items							
P-Value	Algebra		Biology		English		Government	
	N	%	N	%	N	%	N	%
$P < 0.10$	0	0.00	0	0.00	0	0.00	0	0.00
$0.10 \leq P < 0.20$	1	0.67	2	1.46	1	0.87	0	0.00
$0.20 \leq P < 0.30$	11	7.33	9	6.57	2	1.74	7	4.67
$0.30 \leq P < 0.40$	29	19.33	32	23.36	24	20.87	34	22.67
$0.40 \leq P < 0.50$	45	30.00	35	25.55	27	23.48	48	32.00
$0.50 \leq P < 0.60$	28	18.67	27	19.71	29	25.22	29	19.33
$0.60 \leq P < 0.70$	19	12.67	22	16.06	24	20.87	19	12.67
$0.70 \leq P < 0.80$	10	6.67	10	7.30	7	6.09	10	6.67
$0.80 \leq P < 0.90$	6	4.00	0	0.00	1	0.87	3	2.00
$P \geq 0.90$	1	0.67	0	0.00	0	0.00	0	0.00
Descriptive Statistics								
N Items*	150		137		115		150	
Mean	0.49		0.48		0.51		0.49	
SD	0.16		0.14		0.13		0.13	
Min	0.15		0.16		0.18		0.23	
Max	0.90		0.79		0.81		0.89	

* N Items includes the number of unique items; some Biology and English items appear on both Forms 108 and 208.

Table 3.3 Distributions of Point-Biserial Correlations: May, All Mod-HSA Items – Linking

May 2008	Number and Percentage of Items							
Correlation	Algebra		Biology		English		Government	
	N	%	N	%	N	%	N	%
$R < 0.10$	2	1.33	2	1.46	2	1.74	0	0.00
$0.10 \leq R < 0.20$	11	7.33	9	6.57	0	0.00	3	2.00
$0.20 \leq R < 0.30$	34	22.67	25	18.25	32	27.83	13	8.67
$0.30 \leq R < 0.40$	50	33.33	45	32.85	47	40.87	45	30.00
$0.40 \leq R < 0.50$	45	30.00	51	37.23	32	27.83	75	50.00
$0.50 \leq R < 0.60$	7	4.67	5	3.65	2	1.74	14	9.33
$0.60 \leq R < 0.70$	1	0.67	0	0.00	0	0.00	0	0.00
$R \geq 0.70$	0	0.00	0	0.00	0	0.00	0	0.00
Descriptive Statistics								
N Items*	150		137		115		150	
Mean	0.35		0.36		0.35		0.40	
SD	0.11		0.10		0.08		0.08	
Min	-0.03		0.00		0.07		0.12	
Max	0.61		0.54		0.52		0.58	

* N Items includes the number of unique items; some Biology and English items appear on both Forms 108 and 208.

Table 3.4 Distributions of Point-Biserial Correlations: May, All Mod-HSA Items – Target

May 2008	Number and Percentage of Items							
Correlation	Algebra		Biology		English		Government	
	N	%	N	%	N	%	N	%
$R < 0.10$	11	7.33	17	12.41	8	6.96	8	5.33
$0.10 \leq R < 0.20$	26	17.33	27	19.71	19	16.52	20	13.33
$0.20 \leq R < 0.30$	53	35.33	47	34.31	48	41.74	63	42.00
$0.30 \leq R < 0.40$	48	32.00	39	28.47	34	29.57	54	36.00
$0.40 \leq R < 0.50$	12	8.00	7	5.11	6	5.22	4	2.67
$0.50 \leq R < 0.60$	0	0.00	0	0.00	0	0.00	1	0.67
$0.60 \leq R < 0.70$	0	0.00	0	0.00	0	0.00	0	0.00
$R \geq 0.70$	0	0.00	0	0.00	0	0.00	0	0.00
Descriptive Statistics								
N Items*	150		137		115		150	
Mean	0.26		0.24		0.26		0.27	
SD	0.11		0.11		0.10		0.09	
Min	-0.07		-0.08		-0.05		0.03	
Max	0.48		0.48		0.45		0.50	

* N Items includes the number of unique items; some Biology and English items appear on both Forms 108 and 208.

Table 3.5 May Mod-HSA Items Excluded from Calibration

May 2008	MD ID	Form	Sequence	Response Type	Reason
Content					
Algebra	258175	108	64	M	Rbis=-0.04
Biology	261615	208	31	M	Rbis= 0.00
English	259462	208	47	M	Rbis= 0.08

Table 3.6 May Mod-HSA Items with Statistical Flags Retained in Calibration

	P-Value < 0.10	P-Value > 0.90	R_ITT < 0.10	Distractor Pt-Bis > 0	Omit Rate SR/SPR > 5%	C-Level DIF	Total Flags	N Items
May 2008								
Algebra	0	18	1	0	0	9	28	28
Biology	0	12	1	4	0	1	18	18
English	0	9	1	1	0	2	13	11
Government	0	20	0	0	0	4	24	23

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 that 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 the Mantel-Haenszel DIF detection method to assess differential SR item performance. 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. 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 a lower mean item score than the focal group.

There were 16 items flagged for C-level DIF involving one or more of the identified focal groups (i.e., female, African American, American Indian, Asian, Hispanic). The items flagged for C-category DIF included nine Algebra items, one Biology item, two English items and four Government items. These items were retained in the Mod-HSA item bank, and will be reviewed and evaluated to determine their eligibility for future use.

---

[4] 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, 1988):
$MH\ D\text{-}DIF = -2.35 \ln[\alpha_{MH}]$.