Section 3. Validity

Validity is one of the most important attributes of assessment quality and is a fundamental consideration when tests are developed and evaluated (AERA, APA, & NCME, 1999; Messick, 1989). Validity refers to the degree to which logical, empirical, and judgmental evidence supports each proposed interpretation or use of a set of scores. Validity is not based on a single study or type of study but is an ongoing process of gathering evidence to support the interpretation or use of the resulting test scores. The process begins with the test design and continues throughout the entire assessment process, including content specifications, item development, psychometric quality, and inferences made from the test results.

Students' scores on an MD HSA are inferred to reflect students' level of knowledge and skills in a content area. The scores are used to classify students in terms of their level of proficiency using cut scores established by the state.

Evidence Based on Analyses of Test Content

The MD HSAs are referred to as end-of-course tests because students take each test as they complete the appropriate coursework. Consequently items are developed to measure the knowledge and skills expected of students following completion of coursework. As discussed in Section 2, the development of test content for each MD HSA is overseen by a content expert who has a depth of knowledge and teaching experience related to the course in which the MD HSA is to be administered. Appropriate content leads who have similar qualifications review the test development work of these individuals.

Evidence based on analyses of test content includes logical analyses that determine the degree to which the items in a test represent the content domain that the test is intended to measure (AERA, APA, & NCME, 1999, p. 11). The test development process for the MD HSAs provides numerous opportunities for the MSDE to review test content and make changes to ensure that the items measure the knowledge and skills of Maryland students according to course standards. Every item that is created is referenced to a particular instructional standard (i.e., goal, expectation, or indicator). During the internal ETS development process, the specific reference is confirmed or changed to reflect changes to the item. When the item is sent to a committee of Maryland educators for a content review, the members of the committee make independent judgments about the match of the item content to the standard it is intended to measure and evaluate the appropriateness for the age of students being tested. These judgments are tabulated and reviewed by the content experts, who use the information to decide which items will advance to the field test stage of development.

Evidence Based on Analyses of Internal Test Structure

Analyses of the internal structure of a test typically involve studies of the relationship among test items and/or test components in the interest of establishing the degree to which the items or components appear to reflect the construct on which a test interpretation is based (AERA, APA & NCME, 1999, p. 13). The term construct is used here to refer to the characteristic that a test is

intended to measure; in the case of the MD HSAs, the characteristic of interest is the knowledge and skills defined by the test blueprint for each content area.

These test blueprints are derived from Maryland's Core Learning Goals for each course. The test blueprints are presented in Section 2 (see Tables 2.2 to 2.5); the CLGs can be found on the MSDE website at http://www.mdk12.org/assessments/high_school/index_a.html.

High total group internal consistencies as well as similar reliabilities between subgroups with roughly the same sample size provide additional evidence of validity. High reliability over items within a test implies that the test items within a domain are measuring the intended construct. Cronbach's alpha results for each administration for the overall population, as well as for subgroups can be found in Section 6 of this report, in Tables 6.5 through 6.24.

Another way to assess the internal structure of the test is through the evaluation of Pearson correlation matrices between the individual MD HSA subscores. If subscores within a content area are strongly related to each other, this is another indicator of validity. Tables 3.1 and 3.2 highlight the Pearson correlations found between subscores within each content area of the MD HSA tests. Results indicate that each subscore is significantly positively correlated with the total Scale Score as well as the individual subscores measured in each content area.

Table 3.1 Correlations between Subscores by MD HSA 2011 Content Area – Algebra & Biology

	Algebra					Biology						
	SS	1	2	3	4	SS	1	2	3	4	5	6
Overall SS	1					1						
Subscore 1	.82	1				.78	1					
Subscore 2	.83	.65	1			.73	.57	1				
Subscore 3	.78	.60	.62	1		.77	.57	.57	1			
Subscore 4	.68	.52	.54	.54	1	.74	.57	.54	.54	1		
Subscore 5	-	-	-	-	-	.69	.53	.50	.50	.50	1	
Subscore 6	-	-	-	-	-	.76	.57	.56	.56	.54	.51	1

Note: All correlations significant at the $p < .001$ level.

Table 3.2 Correlations between Subscores by MD HSA 2011 Content Area – English & Government

	English					Government					
	SS	1	2	3	4	SS	1	2	3	4	5
Overall SS	1					1					
Subscore 1	.78	1				.85	1				
Subscore 2	.74	.56	1			.85	.68	1			
Subscore 3	.71	.49	.47	1		.74	.62	.60	1		
Subscore 4	.72	.51	.47	.50	1	.73	.63	.59	.54	1	
Subscore 5	-	-	-	-	-	.72	.58	.59	.54	.51	1

Note: All correlations significant at the $p < .001$ level.

Finally, the internal structure of the MD HSA tests can be assessed in relation to the degree to which these tests meet the requirements of the statistical models used throughout testing administrations. Confirmatory factor analyses (CFAs) for each test by content area can be conducted to examine the underlying domain structure of the MD HSA test. CFA is a useful statistical methodology as it can evaluate whether performance on items in each test reflects a single underlying characteristic or a set of distinct characteristics defined by the reporting categories for each content area. The findings from this type of analysis are helpful as they can establish whether the unidimensional model-based IRT used to calibrate the MD HSA items is appropriate.

Confirmatory Factor Analyses of the May 2011 Administration Data

To assess the dimensionality of the MD HSA tests, CFA's for each content area were conducted using test data from the primary forms of the May 2011 administration. The May administration was chosen for analysis because it is the largest and most representative administration of the MD HSAs. The May administration consisted of ten primary forms; data from operational items were combined across forms within the content areas of Algebra, Biology, English, and Government.

Mplus (Muthén & Muthén, 2007) was used to evaluate unidimensional, or one-factor, CFA models for each content area. As item level data on the MD HSA tests are dichotomous, methods available in *Mplus* that take into account the categorical nature of the data were used (see Muthén, 1998-2004).

Model parameter estimation was accomplished using a weighted least-square method with a mean and variance adjustment (*WLSMV*; Muthén, DuToit, & Spisic, 1997). This method leads to a consistent estimator of the model parameters and provides standard errors that are robust under model misspecification. For categorical data, this estimation method offers an alternative to full-weighted least squares (WLS) techniques that generally become computationally too demanding for models with more than a few observed variables (items).

Overall model fit for each CFA model by content area was examined using the scaled chi-square (χ^2) test of model fit in combination with supplemental fit indices. The Tucker-Lewis Index (TLI) compares the chi-square for the hypothesized model with that of the null or "independence" model, in which all correlations or covariances are zero. TLI values range from zero to 1.0, and values greater than 0.94 signify good fit (Hu & Bentler, 1999). The comparative fit index (CFI) and root mean square error of approximation (RMSEA) index both are based on noncentrality parameters. The CFI compares the covariance matrix predicted by the model with the observed covariance matrix, and the covariance matrix of the null model with the observed covariance matrix. A CFI value greater than 0.90 indicates acceptable model fit (Hu & Bentler, 1999). The RMSEA assesses the error in the hypothesized model predictions; values less than or equal to 0.06 indicate good fit (Hu & Bentler, 1999).

Table 3.3 shows the results of the analyses. The TLI, CFI, and RMSEA fit statistics indicated that the one-factor solutions fit the data well in all content areas. Though none of the χ^2 results indicated good fit using a criterion of $p > .05$, this was expected due to the extremely large sample sizes. These findings provide evidence that the tests for each content area measure a single dimension. This is a positive finding, given that IRT models assume unidimensionality.

Table 3.3 MD HSA 2011 Confirmatory Factor Analyses Fit Statistics

Content	Admin	Forms	# of Factors	# of Items	*n*	*df*	χ^{2*}	TLI	CFI	RMSEA
Algebra	May	D-H, J-N	1	53	70,511	1,325	65,267	**0.97**	**0.97**	**0.026**
Biology	May	D-H, J-N	1	76	55,919	2,774	60,848	**0.98**	**0.98**	**0.019**
English	May	D-H, J-N	1	60	56,107	1,710	35,850	**0.97**	**0.97**	**0.019**
Government	May	D-H, J-N	1	82	53,050	3,239	102,174	**0.95**	**0.95**	**0.024**

Note: Table entries that meet or exceed the criterion are in bold.
* $p < .0005$ for all χ^2

Speededness

If more than five percent of students omitted an SR or SPR item the item was flagged as having a high omit rate. Table 3.4 shows omit rates for each content area by administration and item type. As can be seen, several of the SPR items were flagged for having high omit rates. This pattern is consistent with findings from previous test years. Note that in general, SPR items tend to have higher omit rates because students have to generate a response rather than choose one from the available answer choices. None of the SR items were flagged.

Table 3.4 Number of MD HSA 2011 Operational Items Flagged for High Omit Rate

Content	October		January		April		May		Summer	
	Item Types		Item Types		Item Types		Item Types		Item Types	
	SR	SPR	SR	SPR	SR	SPR	SR	SPR	SR	SPR
Algebra	0	4	0	2	0	8	0	0	0	4
Biology	0	--	0	--	0	--	0	--	0	--
English	0	--	0	--	0	--	0	--	0	--
Government	0	--	0	--	0	--	0	--	--	--

The percentage of students who respond to the last items in a test can be used to assess the degree to which a test is speeded. When speededness occurs, a test is measuring not only students' knowledge and skills as defined by the construct of interest but also the speed at which

the knowledge and skills are demonstrated, which is a second construct. In tests of achievement, it is desirable to find that speededness is not present in a test, which provides evidence that student scores on the test reflect only the intended construct. Evidence of speededness is provided by the finding that the omit rates at the end of a test are notably higher than those observed elsewhere in the test.

Appendix 1.A presents the percentage of students who omitted items on the MD HSA operational forms. Across all content areas and administrations, the percentage of students who did not respond to the last ten SR items of a test was less than 2 percent per item. The SPR items (Algebra only) had omit rates ranging from 0.2 percent to 7.0 percent when placed within the last ten items of a test form. The higher omission rates for the SPR items are typical for this item type because students are required to solve a problem and then record the answer in an answer grid, rather than choose from among four answer choices presented by SR items. For all item types the percentage of students who omitted items located within the last ten positions on an MD HSA test form was not greater than omit rates throughout the test.

In addition to the factor analyses and the information regarding speededness presented here and the validation documentation gathered and maintained by MSDE, other information in support of the uses and interpretations of MD HSA scores appears in the following sections:

- Section 4 provides detailed information concerning the scores that were reported for the MD HSAs and the cut scores for each content area.
- Section 5 provides information concerning the test characteristics based on classical test theory for the administrations of the MD HSAs.
- Section 6 presents information regarding student characteristics for the administrations of the MD HSAs.
- Section 7 includes documentation regarding the field test analyses. Descriptions of classical item analyses, differential item functioning, item response theory calibration, and scaling are included. In addition, summary tables of item *p*-value and item-total correlation distributions are provided.