Section 10. Validity

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

Students' scores on the MD Mod-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, based on cut scores established by the state.

Evidence Based on Analyses of Test Content

The Maryland Mod-HSAs are referred to as end-of-course tests because students take each test as they complete the appropriate coursework. Banked HSA items were selected and adapted for the MD Mod-HSAs to measure the knowledge and skills expected of students following completion of coursework.

The constructs measured by each MD Mod-HSA are described in detail in the Maryland high school curriculum standards, or Core Learning Goals. All ETS content staff working on item selection and development have been trained in the CLGs. The test blueprint documents presented in Section 9 (see Tables 9.1 to 9.3) were created in collaboration with committees of Maryland educators and were derived from the Maryland goals, expectations, and indicators.

The process of selecting and adapting banked MD HSA items for use as MD Mod-HSA items is summarized briefly in Section 9 and described in detail in the Maryland Modified High School Assessment 2008 Technical Report.[10] Banked items were referenced to a particular instructional standard (i.e., goal, expectation, or indicator). During the internal ETS development process, the specific reference was confirmed or changed to reflect changes to the item. When the item went to a committee of Maryland educators for content review, the members of the committee made independent judgments about the match of the item content to the standard it was intended to measure and evaluated the appropriateness for the age and cognitive ability of the students being tested.

[10]Available at http://marylandpublicschools.org/MSDE/divisions/planningresultstest/HSA+Technical+Reports.htm

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 Mod-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 9 (see Tables 9.1 through 9.3); the CLGs can be found on the MSDE website at
http://www.mdk12.org/assessments/high_school/index_a.html

High internal consistency, as is discussed in Section 12, constitutes evidence of validity as high reliability over items within a test implies that the test items within a domain are measuring a single construct. The internal structure of the test can also be assessed through the evaluation of correlation matrices between individual MD Mod-HSA subscores. Subscores that are strongly related to each other are indicators of construct validity. Tables 10.1 and 10.2 contain the Pearson correlations found between subscores within each content area of the MD Mod-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 10.1 Correlations between MD Mod-HSA Subscores by Content Area – Algebra & Biology

	Algebra					Biology						
	SS	1	2	3	4	SS	1	2	3	4	5	6
Overall SS	1					1						
Subscore 1	.63	1				.65	1					
Subscore 2	.58	.35	1			.39	.23	1				
Subscore 3	.76	.34	.32	1		.56	.27	.20	1			
Subscore 4	.39	.25	.27	.26	1	.50	.27	.22	.26	1		
Subscore 5	-	-	-	-	-	.45	.24	.18	.23	.21	1	
Subscore 6	-	-	-	-	-	.49	.24	.21	.27	.25	.22	1

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

Table 10.2 Correlations between MD Mod-HSA Subscores by Content Area – English

	English				
	SS	1	2	3	4
Overall SS	1				
Subscore 1	.63	1			
Subscore 2	.64	.42	1		
Subscore 3	.69	.37	.39	1	
Subscore 4	.58	.29	.27	.33	1

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

Evidence of the internal structure of an assessment also comes from evaluation of the dimensionality of the test — whether performance on items that comprise each test reflects a single underlying characteristic or a set of distinct characteristics. Previous exploratory factor analytic (EFA) studies evaluating the dimensionality of the MD Mod-HSA operational forms have indicated a single underlying dimension[11]. To establish the validity of a single factor model over test content areas on the MD Mod-HSA, confirmatory factor analysis (CFA) can be used.

Confirmatory Factor Analyses of the May 2012 MD Mod-HSA Administration Data

To assess the dimensionality of the MD Mod-HSA tests, CFA's for each content area were conducted using test data from the primary form of the May 2012 administration. The May administration was chosen for analysis because it is the largest and most representative administration of the MD Mod-HSAs. The May administration consisted of one primary form; data from operational items were combined across forms within the content areas of Algebra, Biology, and English.

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 the full-weighted least square (WLS) technique that generally becomes computationally too demanding for models with more than a few observed variables (items).

[11] When the Mod-HSA was created in 2008, exploratory factor analysis (EFA) was conducted. Please refer to the Maryland Modified High School Assessment 2008 Technical Report for details.

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 around 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. Higher CFI values indicate better 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 10.3 shows the results of the analyses. Fit statistics indicated that the one-factor solutions generally fit the data well in all content areas. Although none of the χ^2 results indicated good fit given the criterion of $p > .05$, this was expected because of the large sample sizes. These findings provide evidence suggesting the tests for each content area measure a single underlying dimension. This is a positive finding, given that IRT models assume unidimensionality.

Table 10.3 MD Mod-HSA Confirmatory Factor Analyses Fit Statistics

Content	Forms	# of Factors	# of Items	*n*	*df*	χ^{2*}	TLI	CFI	RMSEA
Algebra	412	1	50	3,491	827	2,673	0.87	**0.84**	**0.025**
Biology	412	1	50	2,369	754	1,172	**0.93**	**0.92**	**0.015**
English	412	1	50	2,526	768	1,303	**0.95**	**0.94**	**0.017**

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

Speededness

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 2A presents the percentage of students who omitted items on the MD Mod-HSA operational forms. The percentage of students who did not respond to the last ten

items of a test was less than 1 percent for all content areas and sessions, with the exception of the Summer forms. One Summer administrations of Algebra (Form 512) had one item with an omit rate of 2.3 percent. The Summer administration of Biology had one omit rate of 1.4 percent. One summer administration of English (Form 612) had omit rates of 2.7 percent for several items. With those exceptions, the item level omit rates for the last ten items are quite low and are comparable to the omit rates for all items. This provides evidence that students had sufficient time to complete the entire test.

Further, if more than 5 percent of students omit a selected response item at any point in the test, the item is flagged as having a high omit rate. No MD Mod-HSA items were flagged for high omit rate in any content area for any administration.

Other information in support of the uses and interpretations of the MD Mod-HSA scores appears in the following sections:

- Section 11 provides detailed information concerning the scores that were reported and the cut scores for each content area.
- Section 12 provides information concerning test characteristics based on classical test theory.
- Section 13 presents information regarding student characteristics for the MD Mod-HSA administrations.