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 2.3 to 2.6); the CLGs can be found on the MSDE website at http://www.mdk12.org/assessments/high_school/index_a.html.

Confirmatory Factor Analyses

ETS conducted confirmatory factor analyses (CFAs) for the MD HSAs in the interest of investigating whether performance on the 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 the analyses also could be used to establish whether the unidimensional model-based IRT used to calibrate the MD HSA items was appropriate.

Confirmatory factor analyses were conducted using test data from the primary forms of the May 2009 administration. The May administration was chosen for analysis because it is the largest and most representative administration of the MD HSAs; this was also the first administration that did not include the administration of BCR and ECR items. The May administration consisted of eleven 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 calculate matrices consisting of tetrachoric correlations between the items included in each analysis. *Mplus* was also used to fit specified factor models to the data. For each CFA, two models initially were fit to the data: a one-factor model and a multifactor model, where the factors were defined by the items in each reporting category. For example, in MD HSA Biology, a six-factor model specified constructs that measured (1) Skills and Processes of Biology, (2) Structure and Function of Biological Molecules, (3) Structure and Function of Cells and Organisms, (4) Inheritance of Traits, (5) Mechanism of Evolutionary Change, and (6) Interdependence of Organisms in the Biosphere. Four-factor models were specified for Algebra and English, and a five-factor model was specified for Government. The subscores within each content area were not assumed to be independent; consequently the covariance matrices of the latent factors were estimated. Listwise deletion of cases was employed for all analyses.

Parameter estimation was accomplished using a weighted least-square method with mean and variance adjustment (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 nominal data, weighted least squares estimation offers an alternative to full-information maximum likelihood techniques. The latter becomes computationally too demanding for models with more than a few dimensions. Model fit can be assessed through the use of a scaled chi-square statistic. However, the degrees of freedom for the reference distribution of this statistic cannot be computed in the standard way. The correct degrees of freedom are in part determined by the data, and hence different degrees of freedom may be obtained when applying the same model to different data (Muthén, 1998–2004, pp. 19–20).

Model-data fit 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). The weighted root mean square residual (WRMR) is a relatively new fit index that is believed to be better suited to data that include categorical variables; good model fit is indicated by values less than 0.90 (Finney & DiStefano, 2006).

To evaluate model fit, the one-factor and multifactor fit statistics may be compared. In general, if fit statistics are adequate for the one-factor model and improvement in fit statistics is small for the multifactor model, the results suggest that the data are essentially unidimensional.

In the analysis, the input tetrachoric correlation matrix was used to estimate the factor loadings between the indicators (items) and the latent factors (subscores). Also estimated were the correlations between the latent factors, the assumption being that the subscores are related. The collection of estimated correlations between the latent factors is referred to as the psi matrix.

The multifactor models for Biology and English resulted in the estimation of nonpositive definite psi matrices. This finding is due to linear dependencies between two or more latent factors as well as correlations of 1.0 or greater between some of the latent variables within each content area. The occurrence of nonpositive definite psi matrices serves as an indication that the specified factor structure does not adequately fit the data.

Table 3.1 shows the results of the analyses. None of the χ^2 results indicated good fit, given the criterion of $p > .05$; this was expected because the sample sizes were very large. The WRMR did not indicate adequate fit for one-factor or multifactor models for any of the content areas. The remaining fit statistics indicated that the one-factor solutions generally fit the data well in all content areas. These findings provide evidence that the tests for each content area measure a single dimension.

In an effort to overcome the issue of nonpositive definite psi matrices for the Biology and English multifactor models, a second set of analyses was conducted; the results are presented in Table 3.1. For the second set of analyses, the number of factors was reduced for each of the two content areas until the psi matrix was found to be positive definite. For each content area, the two most highly correlated subscores were combined to create a single factor. Subscores 4 and 5 were combined for Biology, while subscores 1 and 2 were combined for English. Combining subscores for these content areas resulted in positive definite psi matrices; however, improvement was not noted in the fit indices. (See Tables 2.4 and 2.5 for descriptions of Biology and English subscores, respectively.)

Table 3.1 MD HSA 2009 Confirmatory Factor Analyses Fit Statistics

Content	Admin	Forms	# of Factors	# of Items	*n*	df	$\chi2^*$	TLI	CFI	RMSEA	WRMR
Algebra	May	C-H, J-N	1	53	57,807	1,179	51,609	**0.98**	**0.94**	**0.027**	5.365
		C-H, J-N	4	53	57,807	1,175	48,963	**0.99**	**0.94**	**0.027**	5.221
Biology	May	C-H, J-N	1	76	55,469	2,466	55,121	**0.99**	**0.94**	**0.020**	4.021
		C-H, J-N	6**	76	55,469	2,455	52,685	**0.99**	**0.94**	**0.019**	3.928
		C-H, J-N	Reduced to 5	76	55,469	2,459	52,760	**0.99**	**0.94**	**0.019**	3.931
English	May	C-H, J-N	1	60	55,557	1,548	36,535	**0.99**	**0.95**	**0.020**	4.062
		C-H, J-N	4**	60	55,557	1,544	33,840	**0.99**	**0.95**	**0.019**	3.907
		C-H, J-N	Reduced to 3	60	55,557	1,546	33,863	**0.99**	**0.95**	**0.019**	3.909
Government	May	C-H, J-N	1	82	55,040	2,754	64,538	**0.99**	**0.95**	**0.020**	4.060
		C-H, J-N	5	82	55,040	2,746	63,981	**0.99**	**0.95**	**0.020**	4.042

Note: Table entries that meet or exceed the criterion are in bold.

* $p < .0005$.

** Indicates the multifactor CFA psi covariance matrix was not positive definite, signifying that at least one latent variable was a linear combination of the other latent variables representing subscores.

Speededness

If more than 5 percent of students omitted an SR or SPR item, or more than 15 percent of students omitted a CR item, the item was flagged as having a high omit rate. Table 3.2 shows omit rates for each content area by administration and item type. Relatively few SR items were flagged for omit rate. Most of the items flagged for high omit rate were SPR and CR items, which tend to have higher omit rates in general because students have to generate a response rather than choose one from the available answer choices. The tendency for SPR and CR items to have higher omit rates is consistent with findings from previous test years.

Table 3.2 Number of MD HSA 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	CR	SR	SPR	CR	SR	SPR	CR	SR	SPR	SR	SPR
Algebra	2	5	2	0	7	4	2	6	6	0	6	0	10
Biology	0	--	3	0	--	2	0	--	6	0	--	0	--
English	0	--	0	0	--	0	0	--	2	0	--	0	--
Government	0	--	2	0	--	3	0	--	6	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 items of a test was less than 5 percent. The only exception was for Algebra SPR items, which, when placed within the last ten items of a test form, had omit rates ranging from 5.2 percent to 14.0 percent. 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 items of an MD HSA test form was not greater 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.