Section 2. 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 a proposed interpretation or use of a set of scores. Validity is a fundamental consideration when tests are developed and evaluated (AERA, APA, & NCME, 1999; Messick, 1989). Validity is not based on a single study or type of study, but involves 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 results.

Students' scores on an HSA are inferred to reflect students' level of knowledge and skills in a subject 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 MDHSAs are referred to as "end-of-course" tests because students take each test as they complete the appropriate coursework. Consequently items are developed to reflect the knowledge and skills expected of students following completion of coursework. The development of test content for each MDHSA is overseen by a content expert who has a depth of knowledge and teaching experience related to the course in which the MDHSA 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 HSAs provides numerous opportunities for the client 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 HSAs the characteristic of interest is the knowledge and skills defined by the test blueprint for each subject area.

These test blueprints are derived from Maryland's Core Learning Goals for each course. The test blueprints are presented in Section 1 (see Tables 1.2 to 1.5); the Core Learning Goals can be found on the MSDE website at http://www.mdk12.org/assessments/high_school/index_a.html.

Confirmatory Factor Analyses

ETS carried out confirmatory factor analyses for the 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 subject area. The findings from the analyses also could be used to establish whether the unidimensional model-based IRT used to calibrate the HSA items was appropriate.

Confirmatory factor analyses (CFAs) were conducted using test data from the primary forms of the May administration for the 2007-2008 school year. The May administration was chosen for analysis because it is the largest and most representative administration of the HSAs. The May administration consisted of 10 primary forms; data from operational items were combined across forms within the content areas of Algebra, Biology and Government. English forms were configured slightly differently, so that the data could not easily be combined within the May administration. Therefore the two May forms with the largest sample sizes (forms E and F) were analyzed separately.

Mplus (Muthén & Muthén, 2007) was used to calculate matrices consisting of polychoric 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 multi-factor model, where the factors were defined by the items in each reporting category. For example in MDHSA Biology, a six-factor model specified constructs measuring: 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 ordinal 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, p. 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) index compares the chi-square for the hypothesized model to that of the null or "independence" model, in which all correlations or covariances are zero. TLI values range from zero to 1.0; values greater than .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 non-centrality parameters. The CFI compares the covariance matrix predicted by the model to the observed covariance matrix and the covariance matrix of the null model to the observed. A CFI value greater than .90 indicates acceptable model fit (Hu & Bentler, 1999). The RMSEA assesses the error in the hypothesized model predictions; values less than or equal to .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 includes categorical variables; good model fit is indicated by values less than 0.90 (Finney & DiStefano, 2006).

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

In the analysis, the input polychoric 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 multi-factor models for each content area resulted in the estimation of non-positive 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 non-positive definite psi matrices serves as an indication that the specified factor structure does not adequately fit the data.

Table 2.1 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 sample sizes were very large. The

WRMR did not indicate adequate fit for one-factor or multi-factor models for any of the content areas. The remaining fit statistics indicate that the one-factor solutions generally fit the data well in all subject areas. The one-factor CFI results for the English test forms were marginal but the findings based on the other fit indices indicated good fit with this model. These findings provide evidence that the tests for each content area measure a single dimension.

In an effort to overcome the issue of non-positive definite psi matrices for multi-factor models, a second set of analyses were conducted; the results are presented in Table 2.1. For the second set of analyses the number of factors was reduced for each content area until the psi matrix was found to be positive definite. For Algebra, Biology and English, the two most highly correlated subscores were combined to create a single factor. Subscores 1 and 2 were combined for both Algebra and English, while subscores 1 and 5 were combined for Biology. Combining subscores for these content areas resulted in positive definite psi matrices; however improvement was not noted in the fit indices. For Government it was necessary to combine the three most highly correlated subscores (subscores 1, 3 and 4) before a positive definite psi matrix was achieved. As with the other content areas, no improvement was observed among the fit statistics. (See Tables 1.2 – 1.5 for descriptions of subscores by content.)

Table 2.1 Confirmatory Factor Analyses Fit Statistics

Subject	Admin	Forms	# of Factors	# of Items	*n*	df	$\chi2^*$	TLI	CFI	RMSEA	WRMR
Algebra	May	D-H, J-N	1	36†	53,393	549	15,730	**0.99**	**0.97**	**0.023**	4.097
		D-H, J-N	4**	36†	53,393	544	13,914	**0.99**	**0.97**	**0.021**	3.849
		D-H, J-N	Reduced to 3	36†	53,393	546	13,942	**0.99**	**0.97**	**0.021**	3.856
Biology	May	D-H, J-N	1	55	54,982	1,259	43,939	**0.99**	**0.95**	**0.025**	4.773
		D-H, J-N	6**	55	54,982	1,249	40,044	**0.99**	**0.95**	**0.024**	4.549
		D-H, J-N	Reduced to 5	55	54,982	1,253	40,158	**0.99**	**0.95**	**0.024**	4.557
English	May	E	1	50	6,491	847	9,167	**0.97**	0.89	**0.039**	2.681
		E	4**	50	6,491	846	7,948	**0.97**	0.90	**0.036**	2.491
		E	Reduced to 3	50	6,491	847	8,011	**0.97**	**0.90**	**0.036**	2.503
		F	1	50	5,708	821	7,626	**0.96**	0.88	**0.038**	2.505
		F	4**	50	5,708	820	6,570	**0.97**	0.90	**0.035**	2.321
		F	Reduced to 3	50	5,708	821	6,614	**0.97**	0.90	**0.035**	2.330
Government	May	D-H, J-N	1	58	56,536	1,351	64,794	**0.99**	**0.94**	**0.029**	5.518
		D-H, J-N	5**	58	56,536	1,344	63,689	**0.99**	**0.94**	**0.029**	5.469
		D-H, J-N	Reduced to 3	58	56,536	1,349	64,299	**0.99**	**0.94**	**0.029**	5.496

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

* $p < .0005$.

† During the May administration two Algebra items were excluded from scoring due to printing errors.

** Indicates the multi-factor 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.

In addition to the factor analyses presented here and the validation documentation gathered and maintained by MSDE, other information in support of the MDHSAs appears in the following sections.

- Section 3 provides detailed information concerning the scores that were reported for the MDHSAs and the cut scores for each content area.

- Section 4 provides demographic information for the population of students who were administered the MDHSAs. Summary statistics at the test level are reported for the student population and for subgroups. In addition, score reliability analyses and measures of decision accuracy and consistency are provided for the student population.

- Section 5 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.