Appendix 1C. Study of the Comparability of Online and Paper Forms of the May 2009 Maryland High School Assessments

As Submitted to the Maryland State Department of Education, November 9, 2009

This memorandum summarizes the results of the May 2009 MD HSA modality comparability study. The analyses as defined by the National Psychometrics Council (NPC) and described in the memorandum ETS provided to MSDE (ETS, personal communication, July 14, 2009) were employed to assess the comparability between the paper and online forms. Revisions requested by the NPC on September 21, 2009, have been incorporated into this version.

Specifically, this study addressed the following two questions:

1. Is the construct invariant between the two modes of test administration?

2. Given that the construct remains the same, is student performance (such as mean, median, various quartiles) similar between the two modes?

In the sections below, the May 2009 MD HSA paper and online administrations first are described briefly. This is followed by a description of the examinee samples and test forms selected for the comparability study. The particular analyses to address the two research questions are then described, and the results are presented. Finally, the research findings are discussed.

Online and Paper Administration of May 2009 MD HSAs

The MD HSAs assess four content areas: Algebra, Biology, English, and Government. A total of 11 primary test forms were administered in May 2009. These forms had common operational items (referred to as primary operational test Form C) and different field test items. Two makeup forms, X and Y, also were administered. Forms X and Y shared at least 80 percent of their operational items with Form C.

All test forms, Forms C, X, and Y, were administered in both the paper and online formats. For the paper tests, the 11 primary test forms were administered during the primary testing week (Week 1). Form X was administered during the first make-up week, and Form Y was administered during the second make-up week. For online tests, the 13 test forms were spiraled equally throughout the three-week testing window. Therefore, in each content area the majority of both online and paper test takers were administered the primary operational test form, Form C.

Test Forms and Student Samples

The analyses were carried out using data from students that took the online (ONL) and paper-and-pencil (PNP) versions of primary Form C in each content area. Decisions about administration mode were made at the school level. Student assignment to the test modes was not random.

The number of items, raw score points, and subscores in Form C for each content area is provided in Table 1. All items were multiple-choice (selected response; SR) except for ten items in algebra that were gridded, called student produced response (SPR) items. All items were dichotomously scored. Raw total scores and subscores are converted to scale scores using item pattern scoring for reporting purposes. The reporting scores are scale scores ranging from 240 to 650.

Table 1 Number of Items and Score Points in Form C for Each Content Area

Content	No. selected response (SR) items	No. student produced response (SPR) items	No. total items	Possible total raw score points	No. of subscores
Algebra	43	10	53	53	4
Biology	76	-	76	76	6
English	60	-	60	60	4
Government	82	-	82	82	5

Students meeting any of the following criteria were excluded from the analyses: (a) test record invalidated by the test administrator, (b) incorrect form code, or (c) no responses to the first 5 items. Table 2 provides the student sample sizes by test mode and content area.

Table 2 Test Score Summary by Content Area and Test Mode

Content	Test mode	Sample size	Raw scores		Scale scores		Cronbach's alpha
			Mean	SD	Mean	SD	
Algebra	Online	10,888	33	10.8	433	40.5	0.92
	Paper	66,083	30	10.9	423	42.3	0.92
Biology	Online	7,004	45	14.1	430	38.1	0.93
	Paper	49,831	40	14.1	416	43.1	0.93
English	Online	7,196	43	10.4	416	31.4	0.91
	Paper	49,292	40	11.3	407	34.7	0.92
Government	Online	7,268	53	14.6	428	36.6	0.93
	Paper	48,729	47	15.6	414	40.5	0.94

Analyses Pertaining to Construct Invariance

The following analyses were designed to assess whether the same construct was measured by the online and paper versions of the primary operational test administered in each of the four content areas. These analyses focused on the internal structure of the test versions and the degree to which the structures were similar. As noted in the *Standards for Educational and Psychological Testing* (APA, AERA, & NCME, 1999, p. 13), “Analysis of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based.”

Z-score Comparisons

Summary statistics obtained for the items administered in each mode were calculated. Percent correct values (p-values) for the items were converted to z-scores and plotted to examine the consistency of the items' relative difficulties across the online and paper test modes. Z-scores were calculated using the following formula:

$$z_{im} = \frac{p_{im} - \overline{p}_m}{s_{pm}} \tag{1}$$

where, p_{im} is the p-value for item i within a test mode m, $\overline{p}_m$ is the mean of the items in test mode m, and s_{pm} is the standard deviation of the p-values of the items in test mode m.

In addition, a first principal axis was fit to the scatterplot of z-scores from the two modes for each content area. The first principal axis is the line that minimizes the sum of the squared orthogonal distances between the data points and the line (Niklas, 1994, pp. 328–334). A program called *SMATR* was used to generate the first principal axis in each plot (Falster, Warton, & Wright, 2006). Finally, correlations between the ONL and PNP z-scores were calculated.

Summary Statistics

Table 2 shows the means and standard deviations of total test raw scores and scale scores as well as reliability coefficients (Cronbach's alpha) by content area and test mode.

The students taking the online tests performed better than students taking the paper tests across all content areas. The reliability coefficients were the same or nearly the same across test modes for all content areas; they ranged from 0.91 to 0.94.

Z-Score Comparisons Results

Table 3 shows the item p-value summary by content area and test mode. Items appear to be easier in the online format, as would be expected given the higher total raw scores obtained by the online group.

Table 3 Summary Statistics Describing Item P-values by Content Area and Test Mode

Content	Test mode	No. items	Min	Max	Mean	SD	Median
Algebra	Online	53	0.27	0.88	0.62	0.17	0.66
	Paper	53	0.21	0.85	0.56	0.17	0.62
Biology	Online	76	0.26	0.90	0.59	0.15	0.59
	Paper	76	0.23	0.85	0.52	0.15	0.51
English	Online	60	0.43	0.93	0.72	0.13	0.74
	Paper	60	0.39	0.90	0.66	0.13	0.67
Government	Online	82	0.25	0.97	0.64	0.17	0.67
	Paper	82	0.23	0.95	0.58	0.17	0.59

Figures 1 through 4 contain scatterplots of the item z-scores from both testing modes for the four content areas. Each figure includes the first principal axis. The figures show that in all content areas the data points were very close to the first principal axis. The slopes of the first principal axes are one and the intercepts are zero. There are no outliers in the plots, and correlations between ONL and PNP z-scores ranged from 0.98 to 0.99.

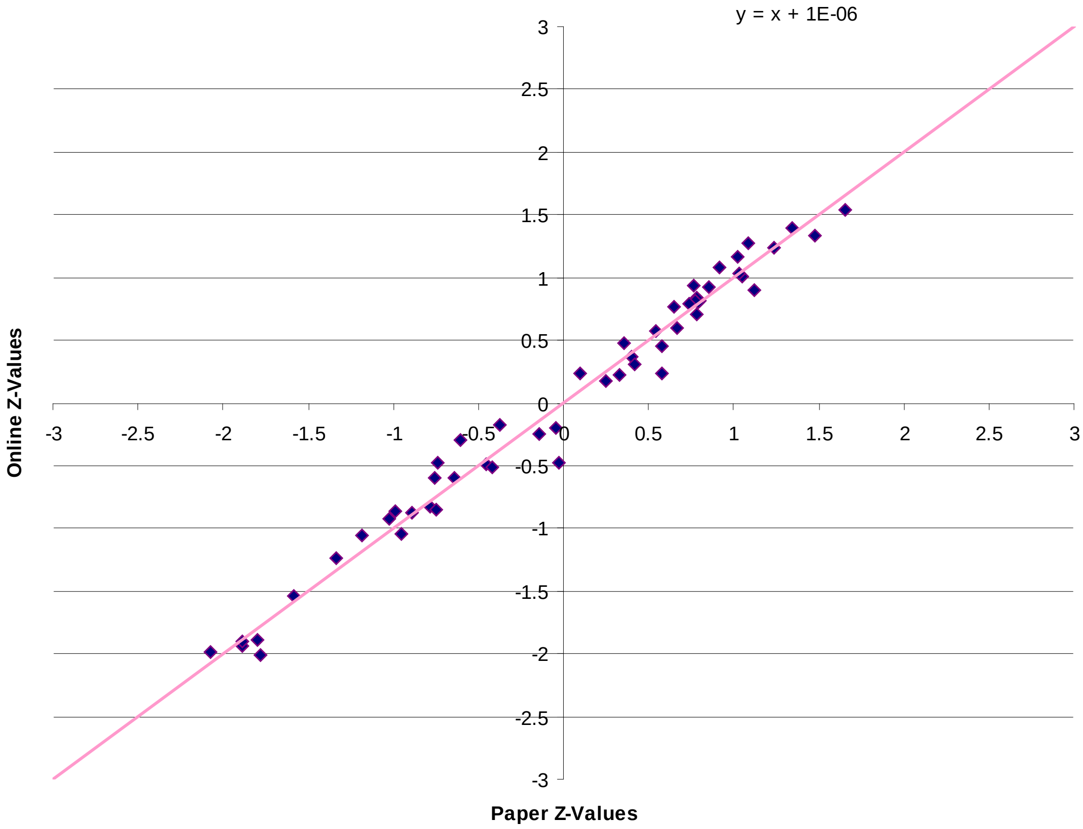

Figure 1. May 2009 HSA—Algebra online and paper z-values and the first principal axis

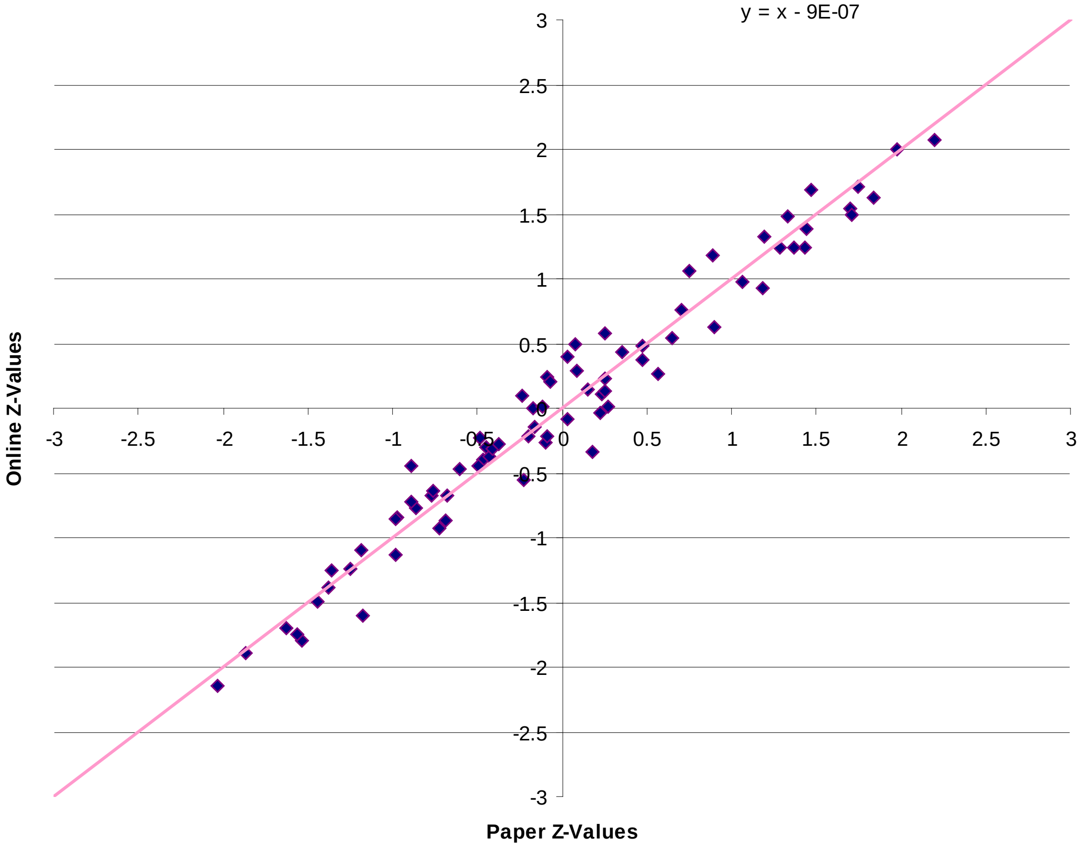

Figure 2 May 2009 HSA—Biology online and paper z-values and the first principal axis

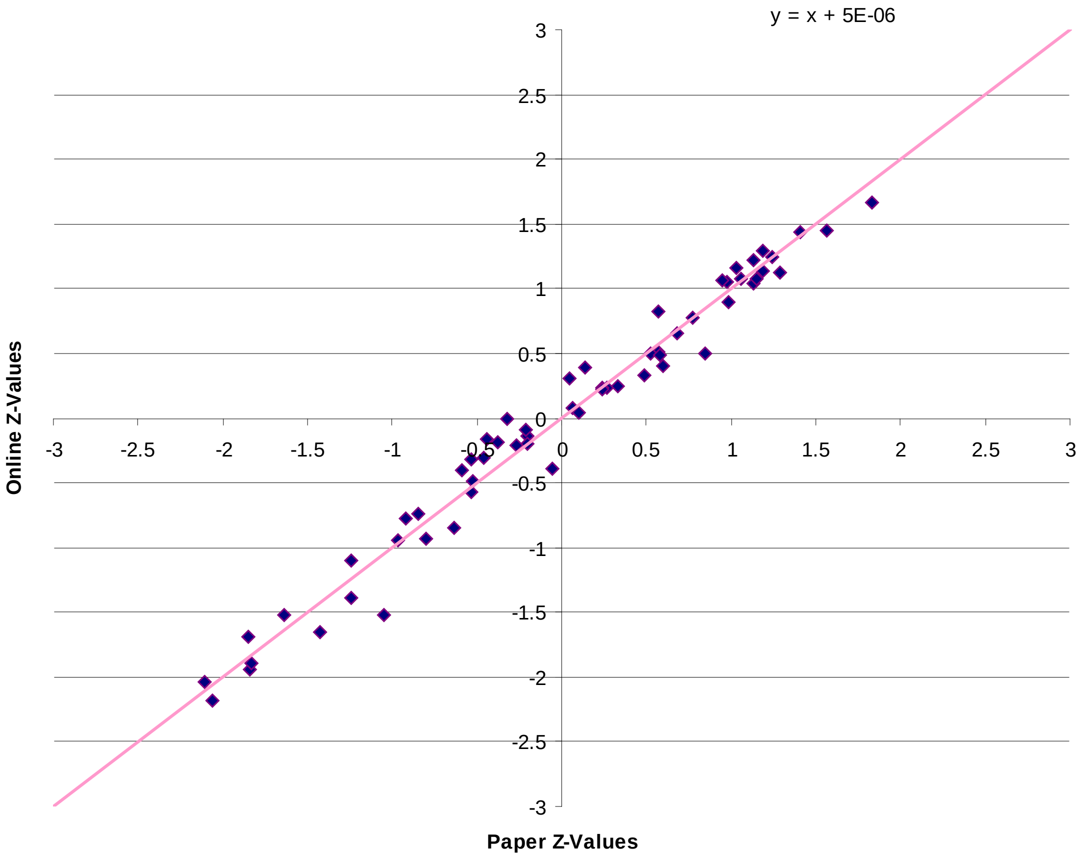

Figure 3 May 2009 HSA—English online and paper z-values and the first principal axis

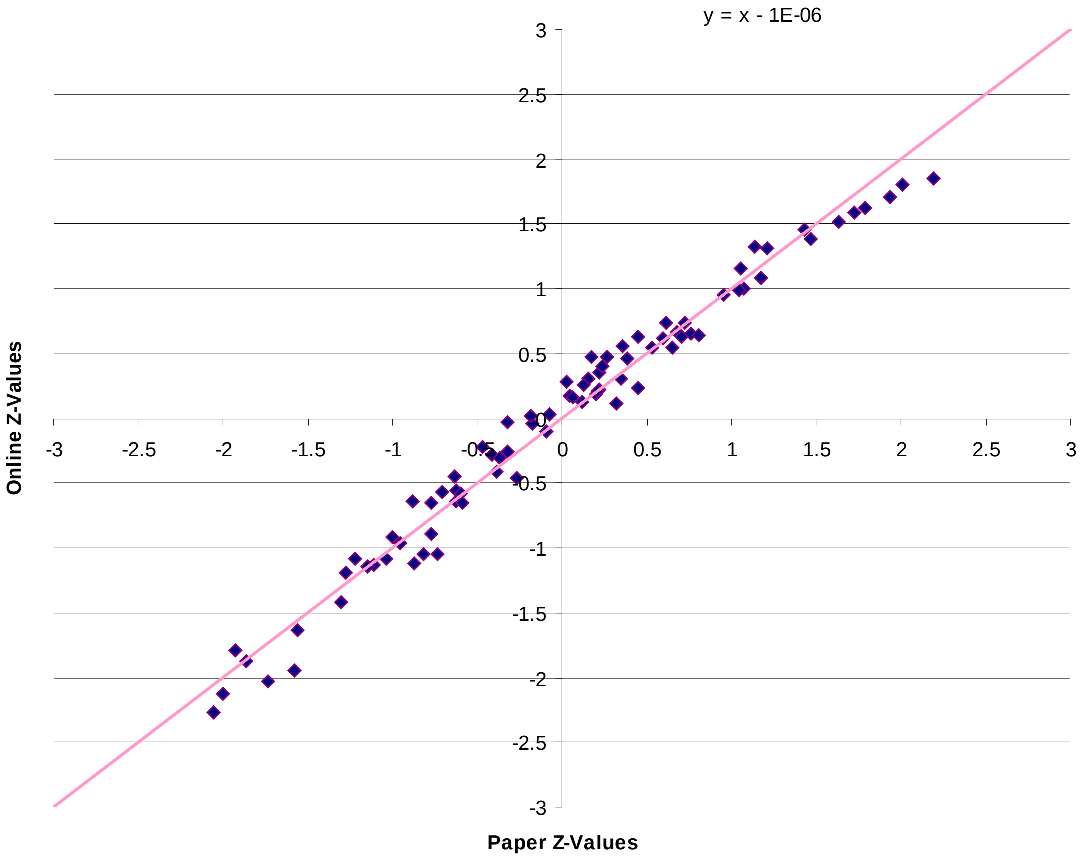

Figure 4 May 2009 HSA—Government online and paper z-values and the first principal axis

Differential Item Functioning

Analyses were carried out to assess differential item functioning (DIF) between the two test modes using the Mantel-Haenszel DIF procedure (MH DIF; Dorans & Holland, 1993; Mantel & Haenszel, 1959). Three DIF analyses were conducted.

The first analysis used students' raw scores as their ability estimates. This is the usual method used to conduct MH DIF analyses. In the second and third analyses an adjustment was made to students' ability estimates to simulate a "small effect size" (SmES) due to administration mode. More specifically, in the second analysis a constant reflecting a small effect size was added to the raw scores of students who took the paper form. In the third analysis, the constant reflecting the small effect size was subtracted from their scores. The constants used to make the adjustments were derived using Cohen's (1988, p. 25) definition of a small effect size:

$$\text{SmES} = 0.2\sqrt{(\sigma^2_{online} + \sigma^2_{paper})/2}\ , \qquad (2)$$

where σ^2_{online} and σ^2_{paper} are the variances of students' total raw scores on the online and paper tests, respectively. The SmESs found for the May 2009 Algebra, Biology, English, and Government HSAs were 2, 3, 2, and 3 (rounded to integer values), respectively.

The logic of assessing DIF using adjusted scores in addition to the unadjusted scores was as follows. A Mantel-Haenszel DIF analysis entails comparing the item performance of two groups of examinees after these examinees have been stratified by ability. Ability is usually measured by the examinees' total test raw scores; students with the same score are grouped together and assumed to be equal in ability. The adjustments were designed to address the possibility that a given total score on the paper test and on the online test did not reflect the same level of ability. It may be, for example, that students taking the paper test got slightly lower scores than did their equally able counterparts who took the online test. Or students who tested on paper might have gotten slightly higher scores than their online counterparts. The purpose of adjusting students' paper scores by adding and subtracting one SmES was to adjust for these kinds of negative or positive mode effects prior to conducting the DIF analyses. If all three results agree, the result would be more robust.

The Mantel-Haenszel procedure was used to classify the three DIF categories as defined in Table 4. Consistent with current ETS practice, only Category C DIF is considered to be a potential threat to item fairness and to warrant further investigation (Educational Testing Service, 2002).

Table 4 Categories of Differential Item Functioning

DIF Category	Definition[a]
A (negligible)	MH D-DIF not significantly different from zero, or has an absolute value smaller than 1.
B (slight to moderate)	1. MH D-DIF in absolute value is significantly different from zero but not from one, and is at least one; OR 2. MH D-DIF in absolute value is significantly different from one, but is smaller than 1.5. Positive values are classified as "B+" and negative values as "B-".
C (moderate to large)	MH D-DIF in absolute value is significantly different from one, and is at least 1.5. Positive values are classified as "C+" and negative values as "C-".

Note. [a] the significance level at 0.05.

Differential Item Functioning

DIF classifications by content area for all three DIF analyses are given in Table 5. Results of the DIF analyses showed that no item was found to have C-level DIF in any of the three DIF analyses, and only one item was found to have B-level DIF.

Table 5 DIF Categorization of Items by Content and Type of DIF Analysis

Content area	DIF Category	Raw Score	+ 1 Small Effect Size	- 1 Small Effect Size
Algebra	A	53	53	53
	B	0	0	0
	C	0	0	0
Biology	A	75	76	75
	B	1	0	1
	C	0	0	0
English	A	60	60	60
	B	0	0	0
	C	0	0	0
Government	A	82	82	82
	B	0	0	0
	C	0	0	0

Confirmatory Factor Analyses

Confirmatory factor analyses (CFAs) were carried out in each content area to examine the consistency of subscore structures across test administration modes. The MD HSA blueprints define a subscore structure for each content area.

The first set of CFAs was conducted using item level data. These analyses were designed to investigate the question of whether the subscore structures were the same in the tests administered in the paper and online modes. Table 6 shows the number of items in each content area subscore.

The second set of CFAs was conducted using subscore level scale scores to assess the structural invariance of the paper and online tests. In addition to fitting a single factor model to each test, the fit of three multigroup CFA models that differed in their structural constraints was analyzed. In Model 1 the subscores of the students taking the paper and online tests were pooled and a single factor model was fit to the data without constraints on the factor loadings or error variances. In Model 2 the factor loadings for the corresponding subscores of the paper and online tests were constrained to be equal across testing modes. In Model 3 the factor loadings as well as the error variances were constrained to be equal for the corresponding subscores across testing modes. A comparison of fit results across the three models would demonstrate the degree to which the structure underlying the paper test scores matched the structure underlying the online test scores.

Table 6 Subscore Structures of the May 2009 HSAs

Content area	Subscore Description	No. of items
Algebra	Analysis of patterns and functional relationships	13
	Modeling and interpretation of real-world situations	17
	Collection, organization, analysis and presentation of data	12
	Application of basic concepts of statistics and probability	11
Biology	Skills and processes of Biology	16
	Structure and function of biological molecules	12
	Structure and function of cells and organisms	13
	Inheritance of traits	13
	Mechanism of evolutionary change	9
	Interdependence of organisms in the biosphere	13
English	Reading and Literature: Comprehension and interpretation	16
	Reading and Literature: Making connections and evaluation	14
	Writing: Composing	16
	Language Usage and Conventions	14
Government	U.S. Government Structure, Functions and Principles	23
	Protecting Rights and Maintaining Order	21
	Systems of Government and U.S. Foreign Policy	12
	Impact of Geography on Governmental Policy	11
	Economic Principles, Institutions and Processes	15

All CFAs were conducted using *MPlus* (Muthén & Muthén, 2007). Parameter estimation for the item-level analyses was performed using a weighted least-squares method with mean and variance adjustment (WLSMV; Muthén, DuToit, & Spisic, 1997). This method provides optimal solutions for the analysis of ordered categorical data. The observed variables are binary item responses and, consequently, tetrachoric matrices were used as input for the CFA analyses.

In the item level CFA model, the observed variables (binary item responses) were classified as endogenous dependent variables and the latent factors (i.e., subscores) were classified as exogenous independent variables. In order to scale the factors, the variances of the latent variables were fixed to 1.0. All factor loading patterns were determined based on the defined subscore structures, and factor correlations were freely estimated under the assumption that the subscores could be correlated.

In the subscore level CFA models, maximum likelihood estimation was used. Subscores in the scale score metric were classified as the dependent variables and the latent factors (i.e., total scores) were classified as the independent variables.

Model-data fit was examined using the following 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 0.0 to 1.0; 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 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 0.90 indicates acceptable model fit. The RMSEA assesses the error in the hypothesized model predictions; values less than or equal to 0.06 indicate good fit (Hu & Bentler, 1999). Due to the fact that chi-square and chi-square difference statistics are very sensitive to sample size, Cheung and Rensvold (2002) recommend using various goodness-of-fit indexes to test for measurement invariance. They proposed that when changes in CFI values are smaller than or equal to 0.01, that measurement invariance should not be rejected. Change in CFI values (ΔCFI) are presented in Table 9 for the models testing tau-equivalence (Model 2) and parallelism (Model 3).

Confirmatory Factor Analysis Results

Results of the item-level CFAs for the online and paper tests in each content area are listed in Table 7. The indices show that all the CFA models had adequate and comparable model-data fit. Therefore, the test forms administered in the paper and online modes did not differ in terms of their fit to the specified subscore models.

Table 7 Item-Level CFA Analyses: Model Fit Results

Content area/subscore model	Test mode	RMSEA	TLI	CFI
Algebra 4-Factor Model	Online	0.026	0.987	0.949
	Paper	0.027	0.985	0.939
Biology 6-Factor Model	Online	0.020	0.986	0.953
	Paper	0.019	0.986	0.938
English 4-Factor Model	Online	0.018	0.988	0.965
	Paper	0.020	0.987	0.951
Government 5-Factor Model	Online	0.019	0.989	0.961
	Paper	0.021	0.988	0.942

Results of fitting single factor models to the subscores of the online and paper tests are listed in Table 8. The indices show that the model had adequate and comparable model-data fit; only the RMSEA value for the English online test exceeded the criterion value. Therefore, the test forms administered in the paper and online modes did not appear to differ in terms of their fit to the one factor models when subscores were analyzed. It is also interesting to note that nearly all of the TLI and CFI values given in Table 8 were higher than those given in Table 7, which was based on item level factor analyses, suggesting better model-data fit when subscores were analyzed. The RMSE values given in the two tables were mixed, however, with neither table having clearly better results than the other.

Table 8 Results for the Single Factor Model by Content Area and Mode of Administration

Content Area	Mode	RMSEA	TLI	CFI
Algebra	Paper	.017	.999	1.00
	Online	.024	.998	.999
Biology	Paper	.017	.998	.999
	Online	.027	.996	.997
English	Paper	.050	.987	.996
	Online	.087	.953	.984
Government	Paper	.022	.998	.999
	Online	.020	.999	.999

Results of the series of invariance tests of the one-factor model and fit indices are summarized in Table 9. However, none of the ΔCFI values were greater than .01, suggesting that equivalence constraints on factor loadings and error variances did not reduce model fit. In addition, all of the values for the RMSEA, TLI, and CFI indices exceeded the criteria for good fit for all models. The results differed little over models, suggesting that the construct assessed by the paper and online tests did not differ over modes.

Table 9 Fit Results for Models of Structural Invariance by Content Area

Content Area	Model	RMSEA	TLI	CFI	ΔCFI
Algebra	1	.016	.999	.999	--
	2	.015	.999	.999	0.0
	3	.019	.999	.998	-0.001
Biology	1	.024	.996	.997	--
	2	.022	.997	.997	0.0
	3	.021	.997	.997	0.0
English	1	.045	.989	.994	--
	2	.037	.993	.994	0.0
	3	.040	.991	.990	-0.004
Government	1	.018	.999	.999	--
	2	.017	.999	.999	0.0
	3	.019	.999	.999	0.0

Analyses Pertaining to the Similarity of Student Performance across Modes

Comparisons between test performance of students at selected schools were used to examine whether student performance was similar across groups assessed using different test modes. These comparisons considered both effect sizes and passing rates.

The two May 2009 student groups of interest, those taking the assessments online and those taking the paper-and-pencil assessments were not known to be equivalent because random assignment of students to testing mode was not possible. Consequently, making a direct comparison of the performance of the two groups to assess mode effects on student performance was not appropriate. Therefore, in order to study the comparability of student performance across modes, analyses were conducted at the school level on mean MD HSA performance of matched pairs of schools. In May 2009 schools that tested exclusively in only one mode, either online

(ONL) or paper-and-pencil (PNP) were identified. The reason for using only schools that had tested entirely within a single mode was to minimize any self-selection effects. For each ONL school a matching PNP school was identified. These matched pairs of schools were used for the analyses in this section.

Selection of Schools. First, schools that tested all of their students online in each content area were identified. For each ONL school a matching PNP school was found from among all schools testing their students in only the paper modality.

The main matching variables were schools' May 2007 MD HSA scale score means and standard deviations. The May 2007 scores were chosen as the matching variable because the May 2008 and May 2009 scores were being used in calculation of effect sizes. The small sample of schools to choose from did not allow for school demographic variables to also be considered when matching schools.

The specific steps in the matching process carried out for each content area were as follows:

1. An ONL school was excluded from the matching process and subsequent analyses if it had fewer than 30 students that took either the May 2007, May 2008, or May 2009 test. The numbers of schools excluded were four for Algebra, none for Biology, two for English, and three for Government.

2. For each remaining ONL school, matching PNP candidate schools were identified as those with at least 30 students that took the May 2007, May 2008, and May 2009 tests. Matching PNP candidates also needed to have mean scale score differences from the ONL school of less than one scale score point. If there was no such PNP candidate school, the PNP school having the closest May 2007 mean scale score served as the matching school. There were only a few schools that did not match within one scale score point: three in Algebra, with closest matches of 1.2, 1.4 and 1.9 scale score points, one in Biology that matched by 1.1 scale score points, one in English that matched by 1.8 scale score points, and one in Government with the closest match at 3.0 scale score points.

3. For each ONL school that had more than one potential matching PNP school, the selection criteria were expanded. The magnitude of the difference in test scale score means and standard deviations between May 2007 PNP schools and the ONL schools was considered. Only one PNP school was matched to each ONL school. The resulting numbers of matching pairs of schools were 53 for Algebra, 16 for Biology, 13 for English, and 9 for Government. Addendum A lists the matched pairs by school name.

Calculation of Effect Size. Two effect sizes were calculated for each matched pair of schools. The first effect size was for the May 2008 performance. This effect size was calculated to determine the degree of difference between the groups when all students tested in the paper-and-pencil mode. In that sense, the May 2008 effect sizes served as a baseline for how much of a difference might be expected for reasons other than testing mode.

The second effect size calculated for each pair was based on May 2009 data, when the groups differed by testing mode. If the effect sizes for the May 2009 data were found to be about the same as those for May 2008, this would support the hypothesis that testing mode does not significantly impact overall performance differences. For each pair of schools, the May 2008 and May 2009 effect sizes, $d_{v,}$ were computed as follows (Cohen, 1988, p. 44):

$$d_{ty} = (M_{toy} - M_{tpy}) / \sigma_{ty,pooled} \text{ and,} \tag{3}$$

$$\sigma_{ty,pooled} = \sqrt{(\sigma_{toy}^{2} + \sigma_{tpy}^{2}) / 2} \tag{4}$$

where,

d_{ty} is the effect size in year y (y =2008 or 2009) for school pair t,

M_{toy} and σ_{toy}^{2} are the mean and variance, respectively, of HSA scores of the 2009 ONL school in school pair t in year y,

M_{tpy} and σ_{toy}^{2} are the mean and variance, respectively, of HSA scores of the 2009 PNP school in school pair t in year y, and

$\sigma_{ty,pooled}$ is the pooled standard deviation of the HSA scores in school pair t in year y.

For example, in Algebra, 53 matched pairs of ONL and PNP schools were identified. Therefore, 53 May 2008 effect sizes and 53 May 2009 effect sizes were calculated. A paired t-test was employed to assess whether the average effect size for the 53 ONL and PNP pairs in May 2009 was significantly different from the average effect size calculated using the May 2008 data, when both groups were administered tests on paper.

Calculation of Passing Rates. Because effect sizes could be influenced by extreme low and high test scores, passing rates for the matched schools also were examined. Passing rates are not influenced by extreme scores. Furthermore, passing rates are of interest to stakeholders, such as parents, teachers, and administrators. Passing rates were defined as the percentage of examinees classified as proficient or advanced.

Effect Size and Passing Rate Comparisons at the School Level

Table 10 shows the results of the t-tests comparing the overall effect sizes calculated using May 2008 and 2009 data. The results indicate that the average effect sizes for the ONL and PNP school pairs were not significantly different in the two years. This means that the degree of difference in HSA performance between the two groups of schools was about the same when all students tested on paper (2008) and when the groups of schools tested in different modes (2009). This was true for all content areas. Summary statistics that describe the matched schools by content area are provided in Addendum B.

Table 10 May 2008 and 2009 Effect Sizes for ONL and PNP School Groups

Content	No. of school pairs	May 2008 effect size		May 2009 effect size		*t* statistic	Probability
		Mean	SD	Mean	SD		
Algebra	53	0.02	0.37	-0.03	0.60	0.12	0.90
Biology	16	-0.03	0.29	0.00	0.37	-0.53	0.60
English	13	-0.12	0.18	-0.08	0.23	-0.84	0.42
Government	9	0.03	0.29	0.10	0.26	-1.08	0.31

Table 11 lists the means and standard deviations of the passing rates in the May 2008 and May 2009 administrations by school group and content area. The table shows that the differences between the passing rates for the ONL and PNP schools differed little in 2008 and 2009. The greatest difference occurred in Government where the passing rate difference was 5.9 percent in 2009. Only nine pairs of schools were included in the analyses in this content area, so these results should be interpreted with caution.

Table 11 May 2008 and 2009 Passing Rates and Mean Difference of Passing Rates between Schools That Tested Exclusively Online (ONL) or Paper-and-Pencil (PNP) in May 2009

Content area	No. of schools	Year	School group[a]	Mean (%)	SD (%)
Algebra	53	2008	ONL	87.5	16.2
			PNP	86.5	18.9
			ONL-PNP	1.0	9.0
		2009	ONL	82.8	23.7
			PNP	84.2	21.9
			ONL-PNP	-1.4	20.7
Biology	16	2008	ONL	84.6	11.4
			PNP	84.9	11.3
			ONL-PNP	-0.3	10.6
		2009	ONL	86.1	11.3
			PNP	83.9	13.6
			ONL-PNP	2.1	9.6
English	13	2008	ONL	75.2	13.0
			PNP	78.4	10.4
			ONL-PNP	-3.1	6.6
		2009	ONL	75.9	12.5
			PNP	77.1	13.0
			ONL-PNP	-1.1	9.6
Government	9	2008	ONL	86.2	9.5
			PNP	83.7	9.8
			ONL-PNP	2.5	4.2
		2009	ONL	85.9	10.1
			PNP	80.0	14.5
			ONL-PNP	5.9	6.7

Note: [a] Recall that all students tested in the paper-and-pencil format in May 2008.
ONL = Online schools where all examinees took the May 2009 HSA content test online.
PNP = Paper-and-pencil schools where all examinees took the May 2009 HSA content test in the paper-and-pencil format.

Conclusions

In considering these findings of this study, it is important to note that data from a single test administration were used to evaluate mode effects. If desired, a replication of this study could be conducted following the May 2010 administration if resources are made available.

The current study was conducted to investigate the extent to which the online and paper forms of the MD HSA can be considered to be comparable. The first question of interest was whether the construct was invariant between the two test modes. The internal consistency of the paper and pencil forms was nearly identical to that of the online forms, as were the z-scores. After conditioning on examinee ability, no items were found to function differently across modes. These findings provide evidence that test mode did not significantly affect item performance.

Finally, confirmatory factor analyses showed that, within each content area, the paper and online test forms shared a common subscore structure as defined in the test blueprint. Structural invariance of the models across modes was also demonstrated. In short, there were no findings that suggested that the items administered on paper assessed a different construct than did the items administered online.

The second question addressed whether student performance was similar across the two modes. Comparisons of mean scores and passing rates for matched schools indicated no notable differences in student performance that could be attributed to test administration mode.

Taken together, these results support the use of computer administration of high school assessments in Maryland as equivalent to the existing paper-and-pencil assessments. Further, the use of paper and pencil derived parameters to link the scales of the computer administered assessments to their paper-and-pencil counterpart scales is also supported.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological tests*. Washington, DC: AERA.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H.Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawerence Erlbaum.

Educational Testing Service. (2002). *ETS standards for quality and fairness.* Princeton, NJ: Educational Testing Service.

Falster, D.S., Warton, D.I., & Wright, I.J. (2006). Standardized major axis tests and routines (SMATR; Version 2.0) [Computer software]. New South Wales, Australia. http://www.bio.mq.edu.au/ecology/SMATR/

Holland, P. W., & Thayer, D. T. (1988). Differential item performances and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6(1),* 1–55.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Muthén B. O., & Muthén, L. K. (2007). Mplus 5 [Computer program]. Los Angeles, CA: Muthén & Muthén.

Niklas, K, J. (1994). *Plant allometry: The scaling of form and process*. Chicago: University of Chicago Press.

Appendix A

Table A1 Online and Paper School Pairs Matched on May 2007 MD HSA Mean Scale Scores and Standard Deviations: Algebra

ONL LEA	ONL School	ONL N	ONL Mean	ONL SD	PNP LEA	PNP School	PNP N	PNP Mean	PNP SD
BALTIMORE	Southwest Academy	160	408.05	27.58	BALTIMORE	Parkville High & Center For Math/S	363	407.38	30.07
BALTIMORE	Woodlawn Middle	119	413.27	24.83	ANNE ARUNDEL	Meade High	362	413.67	34.26
BALTIMORE	Windsor Mill Middle	87	416.34	26.58	WICOMICO	James M. Bennett High	170	416.14	27.37
BALTIMORE	Cockeysville Middle	193	457.67	24.79	BALTIMORE	Franklin Middle	268	457.85	28.06
BALTIMORE	Dumbarton Middle	228	463.73	28.15	ANNE ARUNDEL	Severn River Middle	143	463.85	26.73
BALTIMORE	Loch Raven Technical Academy	96	423.43	21.50	BALTIMORE CITY	Baltimore Freedom Academy	83	423.58	23.13
BALTIMORE	Lansdowne Middle	62	433.79	22.61	HARFORD	Harford Technical High	239	433.72	25.06
BALTIMORE	Middle River Middle	91	443.69	18.45	FREDERICK	Gov. Thomas Johnson Middle	108	443.27	21.11
BALTIMORE	Sparrows Point High	180	405.17	27.25	BALTIMORE CITY	Carver Vocational-Technical High	324	405.31	29.65
CHARLES	Milton M. Somers Middle School	170	470.09	29.58	ANNE ARUNDEL	Crofton Middle	187	471.34	27.52
CHARLES	Piccowaxen Middle School	71	455.58	19.45	HARFORD	North Harford Middle	188	455.64	20.39
CHARLES	Thomas Stone High School	205	410.72	24.63	MONTGOMERY	Watkins Mill High	208	410.61	27.73
CHARLES	John Hanson Middle School	136	454.67	21.79	FREDERICK	Thurmont Middle	91	454.98	21.76
CHARLES	Benjamin Stoddert Middle School	70	450.11	25.54	ANNE ARUNDEL	Macarthur Middle	101	450.24	22.37
CHARLES	Westlake High School	232	415.58	28.71	ANNE ARUNDEL	North High	402	415.36	28.64
CHARLES	Mattawoman Middle School	138	434.65	23.45	CECIL	Rising Sun High	164	434.41	22.20
CHARLES	North Point High School	410	428.14	27.30	MONTGOMERY	Northwest High	231	428.03	27.61
CHARLES	Matthew Henson Middle School	67	463.24	24.85	FREDERICK	Ballenger Creek Middle School	129	463.29	22.15
CHARLES	General Smallwood Middle School	60	455.32	14.84	CARROLL	North Carroll Middle	92	455.38	18.37
CHARLES	Henry E. Lackey High School	256	410.70	29.65	ANNE ARUNDEL	Glen Burnie High	414	410.45	30.49
GARRETT	Northern Middle School	57	466.40	21.50	MONTGOMERY	Cabin John Middle School	292	466.17	22.91
GARRETT	Northern Garrett High School	106	426.70	23.67	FREDERICK	Middletown High	107	426.77	26.09
HARFORD	Bel Air Middle	190	466.40	17.49	ANNE ARUNDEL	Central Middle	149	466.96	22.50
HARFORD	Fallston Middle School	179	468.05	20.49	MONTGOMERY	Robert Frost Middle School	241	468.06	20.81
HOWARD	Bonnie Branch Middle	117	468.81	20.47	MONTGOMERY	Robert Frost Middle School	241	468.06	20.81
HOWARD	Ellicott Mills Middle	135	461.59	25.65	QUEEN ANNE'S	Stevensville Middle School	162	461.38	23.34
HOWARD	Howard High	197	434.79	24.76	CECIL	Rising Sun High	164	434.41	22.20
HOWARD	Patapsco Middle	127	477.13	19.46	MONTGOMERY	Takoma Park Middle School	211	477.31	42.99
HOWARD	Dunloggin Middle	87	471.94	22.29	ANNE ARUNDEL	Crofton Middle	187	471.34	27.52
HOWARD	Centennial High	144	449.49	24.04	MONTGOMERY	Briggs Chaney Middle	154	449.36	26.58
HOWARD	Burleigh Manor Middle School	143	472.50	25.56	ANNE ARUNDEL	Crofton Middle	187	471.34	27.52
HOWARD	Mount View Middle	173	463.12	23.65	MONTGOMERY	North Bethesda Middle	207	463.10	22.63
HOWARD	Glenelg High	119	436.01	19.48	BALTIMORE	Sparrows Point Middle	84	435.73	18.55
HOWARD	Glenwood Middle	122	478.74	28.75	MONTGOMERY	Takoma Park Middle School	211	477.31	42.99
HOWARD	Wilde Lake Middle	87	443.03	26.17	CALVERT	Mill Creek Middle	93	443.01	19.36
HOWARD	Harpers Choice Middle	89	461.10	27.38	MONTGOMERY	Ridgeview Middle	154	460.89	23.66
3OWARD	River Hill High	143	462.29	24.76	MONTGOMERY	Julius West Middle	203	462.58	24.65
HOWARD	Lime Kiln Middle	121	473.23	24.45	ANNE ARUNDEL	Crofton Middle	187	471.34	27.52
HOWARD	Cradlerock School	53	444.06	26.38	MONTGOMERY	Thomas S. Wootton High	127	444.00	27.75
HOWARD	Hammond Middle School	123	465.69	22.00	MONTGOMERY	William H. Farquhar Middle	168	465.40	22.62
HOWARD	Oakland Mills Middle	63	461.76	22.74	MONTGOMERY	John H. Poole Middle	89	461.96	26.64

HOWARD	Patuxent Valley Middle	118	436.59	17.78	MONTGOMERY	Benjamin Banneker Middle	129	436.75	21.84
HOWARD	Murray Hill Middle	101	442.97	27.81	CALVERT	Mill Creek Middle	93	443.01	19.36
MONTGOMERY	Richard Montgomery High	145	425.56	30.64	FREDERICK	Brunswick High	131	425.40	31.05
MONTGOMERY	Rockville High	148	430.93	27.12	HARFORD	Fallston High	306	430.61	29.29
MONTGOMERY	Westland Middle	335	465.58	27.40	MONTGOMERY	William H. Farquhar Middle	168	465.40	22.62
MONTGOMERY	Argyle Middle	98	448.20	21.92	MONTGOMERY	Newport Mill Middle	141	447.82	23.56
PRINCE GEORGE'S	Thurgood Marshall Middle School	34	435.56	21.42	MONTGOMERY	White Oak Middle	164	435.85	23.53

Table A2 Online and Paper School Pairs Matched on May 2007 MD HSA Mean Scale Scores and Standard Deviations: Biology

ONL LEA	ONL School	ONL N	ONL Mean	ONL SD	PNP LEA	PNP School	PNP N	PNP Mean	PNP SD
BALTIMORE	Sparrows Point High	122	410.19	28.39	SOMERSET	Crisfield High	38	409.45	28.77
CHARLES	La Plata High School	353	424.92	32.75	MONTGOMERY	Seneca Valley High	125	425.22	31.53
CHARLES	Westlake High School	292	411.09	30.08	DORCHESTER	Cambridge-South Dorchester High	172	411.46	34.02
CHARLES	North Point High School	478	427.53	24.30	HARFORD	C. Milton Wright High	405	427.30	28.26
CHARLES	Henry E. Lackey High School	324	408.23	29.01	DORCHESTER	North Dorchester High School	125	408.34	32.19
GARRETT	Northern Garrett High School	156	422.79	25.07	ANNE ARUNDEL	Arundel High	525	422.82	26.10
HOWARD	Howard High	402	440.45	27.07	FREDERICK	Urbana High	220	441.01	25.08
HOWARD	Centennial High	369	437.94	25.54	BALTIMORE CITY	Baltimore School For The Arts	90	436.83	30.69
HOWARD	Marriotts Ridge High	298	440.24	24.22	FREDERICK	Urbana High	220	441.01	25.08
HOWARD	Glenelg High	260	434.20	31.98	FREDERICK	Walkersville High	164	434.62	31.06
HOWARD	Atholton High	318	437.03	25.63	CALVERT	Northern High	368	436.08	22.57
HOWARD	Reservoir High	323	423.73	37.00	MONTGOMERY	Seneca Valley High	125	425.22	31.53
HOWARD	Long Reach High	308	422.78	33.48	BALTIMORE	Pikesville High	250	422.01	32.48
MONTGOMERY	Richard Montgomery High	441	438.19	38.03	MONTGOMERY	Bethesda-Chevy Chase High	408	439.35	30.90
MONTGOMERY	Rockville High	346	435.30	27.81	SAINT MARY'S	Leonardtown High	298	434.92	27.04
TALBOT	Easton High	233	422.09	30.87	BALTIMORE	Pikesville High	250	422.01	32.48

Table A3 Online and Paper School Pairs Matched on May 2007 MD HSA Mean Scale Scores and Standard Deviations: English

ONL LEA	ONL School	ONL N	ONL Mean	ONL SD	PNP LEA	PNP School	PNP N	PNP Mean	PNP SD
BALTIMORE	Sparrows Point High	177	417.14	28.92	ANNE ARUNDEL	Arundel High	502	417.92	28.87
CHARLES	Thomas Stone High School	371	412.82	33.34	MONTGOMERY	Paint Branch High	371	414.65	30.10
CHARLES	Westlake High School	305	407.94	26.67	ALLEGANY	Fort Hill High	254	407.01	32.39
CHARLES	North Point High School	495	422.32	27.34	CECIL	Rising Sun High	222	422.00	28.58
CHARLES	Henry E. Lackey High School	312	404.80	31.28	FREDERICK	Frederick High	159	404.69	31.30
GARRETT	Northern Garrett High School	148	418.24	26.57	BALTIMORE	Perry Hall High	490	418.29	26.95
HOWARD	Howard High	342	432.19	30.64	ANNE ARUNDEL	Severna Park High	427	431.98	28.93
HOWARD	Marriotts Ridge High	284	434.39	30.19	MONTGOMERY	Poolesville High	204	433.74	30.27
HOWARD	Glenelg High	279	438.02	28.71	BALTIMORE	Dulaney High	456	437.61	34.23
HOWARD	Long Reach High	304	418.84	33.93	BALTIMORE	Catonsville High	230	418.28	35.24
MONTGOMERY	Richard Montgomery High	422	437.88	48.20	MONTGOMERY	Bethesda-Chevy Chase High	392	437.73	36.99
MONTGOMERY	Rockville High	295	422.71	32.50	MONTGOMERY	James Hubert Blake High	469	422.55	32.84
TALBOT	Easton High	256	416.95	31.57	CARROLL	Francis Scott Key High	192	417.22	30.19

Table A4 Online and Paper School Pairs Matched on May 2007 MD HSA Mean Scale Scores and Standard Deviations: Government

ONL LEA	ONL School	ONL N	ONL Mean	ONL SD	PNP LEA	PNP School	PNP N	PNP Mean	PNP SD
BALTIMORE	Sparrows Point High	221	401.27	30.58	PRINCE GEORGE'S	Parkdale High	419	401.50	32.76
CHARLES	La Plata High School	374	425.15	34.46	CARROLL	Westminster High	223	425.40	32.18
CHARLES	North Point High School	555	427.55	30.84	MONTGOMERY	Clarksburg High	370	428.08	30.15
GARRETT	Northern Garrett High School	282	418.58	29.22	WASHINGTON	South Hagerstown High	120	418.24	29.87
HOWARD	Howard High	348	440.32	36.51	MONTGOMERY	Quince Orchard High	371	440.36	38.55
HOWARD	Long Reach High	309	424.68	51.34	MONTGOMERY	Northwood High School	265	424.72	37.24
HOWARD	Glenelg High	273	447.98	31.00	MONTGOMERY	Poolesville High	205	445.03	32.67
MONTGOMERY	Rockville High	290	437.08	34.85	MONTGOMERY	Northwest High	505	436.28	34.68
TALBOT	Easton High	169	428.67	35.61	ANNE ARUNDEL	Meade High	353	428.30	33.08

Appendix B

Table B1 School Level Means and Standard Deviations of Scale Scores by Test Mode: Algebra (53 Pairs)

		2007			2008			2009		
		Sample Size	Scale Score	SD	Sample Size	Scale Score	SD	Sample Size	Scale Score	SD
Online	Mean	133	446	24	135	444	23	127	444	25
Schools	SD	71	21	3	69	20	5	66	27	7
Paper	Mean	186	446	26	183	444	24	192	445	27
Schools	SD	84	21	5	84	21	4	86	25	7
State	Mean	124	421	30	124	423	27	134	422	30
Overall	SD	103	40	13	108	35	11	125	36	12

Table B2 School Level Means and Standard Deviations of Scale Scores by Test Mode: Biology (16 Pairs)

		2007			2008			2009		
		Sample Size	Scale Score	SD	Sample Size	Scale Score	SD	Sample Size	Scale Score	SD
Online	Mean	314	427	29	326	431	29	277	431	32
Schools	SD	93	11	4	94	14	4	84	15	7
Paper	Mean	236	427	29	261	432	29	278	431	34
Schools	SD	134	11	3	154	12	4	140	18	10
State	Mean	169	396	33	194	403	29	191	402	38
Overall	SD	165	40	13	183	33	10	179	34	13

Table B3 School Level Means and Standard Deviations of Scale Scores by Test Mode: English (13 Pairs)

		2007			2008			2009		
		Sample Size	Scale Score	SD	Sample Size	Scale Score	SD	Sample Size	Scale Score	SD
Online	Mean	307	422	32	329	416	30	277	414	30
Schools	SD	91	11	6	101	13	3	92	11	5
Paper	Mean	336	422	31	354	419	31	338	417	31
Schools	SD	128	11	3	132	12	3	137	13	4
State	Mean	185	397	32	202	396	30	190	395	32
Overall	SD	165	31	12	178	28	10	169	28	11

Table B4 School Level Means and Standard Deviations of Scale Scores by Test Mode: Government (9 Pairs)

		2007			2008			2009		
		Sample Size	Scale Score	SD	Sample Size	Scale Score	SD	Sample Size	Scale Score	SD
Online	Mean	313	428	35	300	433	34	234	426	30
Schools	SD	109	13	7	131	13	4	92	13	5
Paper	Mean	315	428	33	382	431	35	387	422	33
Schools	SD	120	13	3	149	17	3	127	16	4
State	Mean	189	401	33	205	407	35	193	402	34
Overall	SD	170	36	10	189	36	12	174	30	11