PPENDIX G: COMPARABILITY STUDY OF PAPER AND PENCIL, AND ONLINE ADMINISTRATION OF THE MOD-MSA

Comparison of Paper-Pencil Version with the On-Line Version of the Maryland Modified School Assessment (Mod-MSA) in Reading and Mathematics (Grades 4 and 5)

In recent years, computer based testing in K-12 settings has become popular in consideration of its many advantages. As Way, Davis and Fitzpatrick (2006) point out, these include savings in cost (no printing and shipment of the paper and pencil format test); improvement in test security; flexibility in test administration; and a base for the utilization of technology in presenting innovative item formats and test delivery algorithms. Above all, on-line (OL) administration provides a quick turnaround of results that could be especially helpful to provide timely feedback to students, teachers, and schools. Furthermore, concern about students' limited familiarity with computers now seems to be displaced by students' preference for computer testing vis-à-vis the paper and pencil (P&P) version (see Glassnapp, Poggio, Poggio, & Yang, 2005).

In comparing the results of the P&P and OL version tests, the main consideration has traditionally been the establishment of a common scale so that scores from the two versions are equivalent. This is often done by matching two groups of students on an external criterion and then comparing their performance. Although an external variable for matching test takers may be difficult to obtain, matching students on a viable external criterion has distinct advantages. The method seems preferable to the costly and rigorous efforts necessary to control for fatigue, student motivation, etc., as would be necessary if a single group of students were to take two versions of a test. Furthermore, assigning (or selecting) students to form randomly equivalent groups may not be a plausible solution because of say, limited technology resources (e.g., a lack of computers at certain schools), or small sample sizes across groups of interest. To date, few studies (e.g., Kim, D. H. & Huynh, H., 2008, 2009; Way, et. al., 2006) have utilized an external matching variable in the comparison of P&P and OL tests.

Once an approach for creating equivalent groups is selected and data in comparable format is obtained or created, analysis of the data is completed using statistical methods, such as, Item Response Theory (IRT), Hierarchical Linear Modeling (HLM), Differential Item Functioning (DIF), Multiple Regression (MR), and Analysis of Covariance (ANCOVA). Selection of an appropriate statistical method not only depends on the design of the study, and availability of data (e.g., large n-counts for IRT), but also on the researcher's goals for the research. For example, the intent may be to compare the performances of the *total* tests across the two groups of examination modes, or the interest may lie in the comparing of *each item's* performance on a test across the two examination modes.

It should be noted that there have been some studies that have compared the P&P and OL versions of tests through the use of item-level analysis. For example, a study by Poggio, Glasnapp, Yang, and Poggio (2005) included both HLM and DIF methods in the analysis, but neither method included an external matching variable. The DIF method was based on random assignment of four forms for the two testing-mode comparison. The HLM method, on the other hand, included three level of analysis where a single group with a counterbalance repeated measure (common persons) design was used for within-student effects at Level 1, and between student effects at Level 2. The students were nested within schools (i.e., each school was assigned to a testing mode), and this variable was used at Level 3. Besides Poggio et al.'s (2005) study, methods using item-level analysis have been used by some other investigators to study mode

differences in test administration, e.g., Keng, McClarty, & Davis, 2006, although Keng, et al.'s study did not use DIF or the external variable design in their research.

Way, et al. (2006), on the other hand, used an external criterion as a matching variable in the comparison of P&P and OL versions of a test. The authors compared test performance of Grade 8 students that tested online with groups from the P&P administration after matching them on their previous spring test performance. In this study, the main purpose of the researchers was to adjust student scores to obtain equivalence across mode of test administration.

This study, as in the previous year (2009) where Grades 7 and 8 were analyzed, uses two methods of comparing test-mode effects. It uses an *external* variable both as a matching variable in forming groups for DIF analysis and as a covariate for the ANCOVA.

Purpose of the Study

The basic requirements for Mod-MSA reading and mathematics assessments do not call for an adjustment to student scores based on the testing modes. The desired goal is only to note the extent to which modes of assessment influences student performance at both the total test level and at the item level.

The purpose of this study, therefore, is:

1. to analyze whether the total Mod-MSA P&P version differs substantially from the OL version with respect to student achievement, and
2. to identify those items that favor one testing mode, and provide this information to MSDE so that steps may be taken to eliminate or modify these items in order to eliminate bias (if bias exists) in Mod-MSA operational forms.

Mod-MSA Reading and Mathematics Assessments

In years prior to the first administration of the Mod-MSA Grades 3-5 tests in spring 2010, approximately 95% of the students, except for Grade 3, regardless of their classification, had taken the MSA examination. Grade 3 students had not taken the MSA in 2009 when they would have been in Grade 2 because the MSA examinations are administered to students starting in Grade 3. Therefore, Grade 3 could not be included in this study.

The Mod-MSA assessments in reading and mathematics were designed for students with disabilities who, based on a decision making process undertaken by their Individual Educational Planning (IEP) team, met specific eligibility criteria. The Mod-MSA tests are alternates to the tests in the MSA Program. The Alternate assessments based on modified achievement standards (AA-MAS) are commonly referred to as 2% assessments. They are specified by the guidelines set by the U.S. Department of Education (DOE) on the basis of the U.S. DOE's Final Rule, of April 9, 2007[1]. According to the rule, although states may test more than 2% of the population using the AA-MAS, they may report only 2% as proficient or above proficiency, for Adequate Yearly Progress (AYP) determinations.

The 2010 Mod-MSA reading and mathematics assessments for Grades 4 and 5 are composed of a mixture of items (unaltered MSA items, modified MSA items, and items created specifically for

[1] U.S. DOE's rule published Monday, April 9, 2007, in the Federal Register as "Title I-Improving the Academic Achievement of the Disadvantaged; Individual of Disabilities Education Act, Final Rule."

the Mod-MSA assessments). The different Mod-MSA item types are intended to provide students access to the grade level content standards that incorporates variation in test delivery through a test that is designed to meet the specific learning characteristics of the students in this population. The format includes standard MSA items from the 2009 administration which were modified to allow students in this population greater access to the material. They also include intact MSA items (for reading), and some new items that were created specifically for the 2010 Mod-MSA administration. Other item modifications include, but are not limited to fewer and shorter reading passages, shorter and less complex questions, and test items with fewer item choices. Both the reading and the mathematics tests had more items administered than were required for the final operational test form. Since the newly created and modified items were administered for the first time during the 2010 administration, some of the items produced statistics that were unacceptable to the Data Review Committee (e.g., negative point biserials). The Committee, therefore, eliminated these items with poor statistics from the selection process. Items were then selected from the remaining pool for the final, scored (operational) form.

Both the Mod-MSA reading and mathematics tests contained only dichotomously scored items (i.e., 45 items for reading and 51 items for mathematics for the operational/scored forms). The items used in the administration of the Mod-MSA assessments were based on Maryland's Voluntary State Curriculum (VSCMSC). The test items for the Mod-MSA were aligned to the VSCMSC for the grade being assessed. Despite the similarity between tests (MSA and Mod-MSA), the Mod-MSA tests are considered separate assessments with a unique set of achievement standards (i.e., cut scores). Furthermore, the Mod-MSA was administered as both a P&P and OL version while the MSA was administered only in the P&P format. Local school systems determined which schools would test online based on the availability of computers. In some cases, special education staff worked with individual students to determine the most appropriate assessment mode for a specific student, after he or she was given the opportunity to take the P&P and OL sample test items.

Research Methodology

The Data Set

Students from Grade 4 and Grade 5, who completed the Mod-MSA, were included in this study. The students' participation for the Mod-MSA was determined by their Individual Educational Planning (IEP) teams. The number of students in this population was expected to be moderate.

Since most of the students who completed the 2010 Mod-MSA in Grades 4 and 5 also took the 2009 MSA in the same subject area, their scores on the 2009 MSA could be used as a covariate and also as an external matching variable for the DIF analyses. The 2009 MSA was administered only as a P&P test, which further enhanced the use of these scores as a covariate or a matching variable because the administration mode variables in the Mod-MSA were not affected by test mode (i.e., the MSA scores) in determining future performance or group classification.

Although almost all 2010 Mod-MSA students in the two grades of interest had a corresponding score on the MSA in 2009, there was no guarantee that it would be possible to match the two sets of scores for each student. The best identifier for matching students on the two tests was the unique State ID. This was the primary matching method used for identifying Mod-MSA students'

2009 MSA test scores. This matching criterion was not perfect, and the fact that a few students may not have taken the MSA in 2009, it was expected that the matched sample thus produced would be smaller than the original Mod-MSA student population.

Methods of Analysis

This study used two methods of comparing test-mode effects by using an external variable that served as a covariate for the analysis of covariance (ANCOVA) and as a matching variable in forming groups for DIF analysis.

a. Test Level Analysis:

In order to examine the differences between groups based on mode of administration, a simple straightforward method would be to use the t-test to determine the significance of the mean differences between groups. However, because the use of a covariate would reduce the within group error and thus produced a more sensitive and powerful test (Stevens, 1990), the ANCOVA was used to compare the P&P and OL version of a test. The covariate in this study was the students' performance on the 2009 MSA because students' Mod-MSA test scores were expected to be positively correlated with their scores on the P&P version of the MSA. As such, students' 2009 MSA scores were considered to be one of the predictors of Mod-MSA test scores, provided the hypothesis of no difference between testing modes for the Mod-MSA examinations was tenable.

A primary benefit of using the ANCOVA method was the *partial* equating of the groups that tested across the two different modes of the Mod-MSA by controlling for students' initial differences (i.e., their differences in achievement on an external variable – the MSA that was expected to correlate with the dependent variable, i.e., the Mod-MSA). Using students' 2009 MSA scores as a covariate equalized the groups on *one* factor, the effects of students' prior knowledge in reading and mathematics that could confound the effects of the testing mode. Instead of testing for significance of the difference in means between the two modes of testing, we tested the difference between the adjusted means of the two modes of testing (i.e. the means that were equalized on the covariate).

It should be noted that the correlation between the Mod-MSA and the MSA scores would be underestimated because of the attenuation effects of the Mod-MSA student population. The population of students taking the Mod-MSA is very different from that of the MSA population. It is likely that these students' MSA scores were in a limited range of the MSA scale vis-à-vis these students' Mod-MSA scores. Nonetheless, we expected the correlation between the dependent and the independent variable to exceed 0.30. This is important, as below this correlation threshold, it is unlikely that the addition of the covariate will lead to an appreciable increase in precision (Cohran, 1957; and Feldt, 1958).

Furthermore, it should also be noted that the assignment of students to the mode of administration was not random. In such cases, the ANCOVA (as in most other statistical analyses) has an important limitation (Anderson, 1963; Lord, 1969). There could be various other variables pertaining to non-randomly assigned schools that may be the cause of differences between the two groups. However, within the limitation of such a possibility, the study gives us a picture of the situation as it exists. This, in our opinion, is justifiable since there is no requirement of adjustments to student scores at this particular time.

In using the ANCOVA, three assumptions regarding the regression part of the covariance analysis, besides those associated with the analysis of variance (ANOVA), have to be met. The ANOVA is fairly robust to violation of its assumptions of normality of the distribution of the dependent variable and the equality of population variance in the two groups. As would be expected, the ANCOVA is robust to the assumptions associated with the ANOVA, but it is also robust to the second of the three additional ANCOVA assumptions listed below, i.e.:

1. a linear relationship between the dependent variable (i.e., the scores on the Mod-MSA) and the covariate exists;
2. the covariate (i.e., the scores on the MSA) is measured without error; and
3. the homogeneity in the population of the regression slopes for the two groups classified on the basis of the testing mode administered (i.e., there is no covariate - MSA scores - by testing-mode interaction).

The first and the last of the three assumptions listed above were checked for tenability prior to the ANCOVA analysis to ensure that these were not violated. In the event that the homogeneity of the regression slopes was not met (i.e., an interaction effect between the covariate and the mode of administration existed) then limits on the regions of non-significance on the covariate were to be established by the use of the Johnson-Neyman technique (Pedhazur, 1973).

b. Item Level Analysis:

For the *item-level* analysis, methods relating to DIF were used to assess the performance of mode effects by items. Groups, based on the mode of test administration, were matched on an external variable (i.e., the students' MSA scores) in this analysis.

Determining that an item is biased requires an inference be made, for which DIF is a necessary, but not sufficient condition (Hambleton, Swaminathan, & Rogers, 1991). Thus, DIF is an important piece of evidence to gather when examining the equivalence of a test across administration modes, but this evidence alone is not sufficient to conclude that an item is biased. This analysis will, however, gives test developers a chance to examine administration method (in this study) with respect to items that may exhibit testing-mode bias, and take this opportunity to eliminate or mitigate the effects of items judged to show bias.

DIF analysis (e.g., the contingency table approaches in Camilli and Shepard, 1994) identifies items that do not function equally between matched groups of individuals. The matching is generally based on equivalency of overall performance, and items that do not perform equally among groups of matched individuals are considered to perform differentially. However, the problem with DIF analysis, in the context of the test-mode comparison, is that student groupings created on the basis of their performance on the administered test may confound the equality criterion of these proficiency groups if the testing mode has systematic differences across some or all items. Specifically, the systematic differences across items will contribute to the score on which students are matched to examine mode effects. It is, therefore, useful to match students on a common non-biased platform (i.e., on a "non-biased" external variable/s prior to DIF analysis).

However, the *unbiased* external criterion must be *an a priori predictor of test performance* for matching students. Because of this, it is important to select an external variable that is not only unbiased with respect to the testing mode, but which would also be a significant predictor of

students' performance on the test if, indeed, no test-mode performance-differences exist for the test.

For this study, the Mod-MSA proficiency -groups for the DIF analysis were based on the Mod-MSA students' performance on the corresponding 2009 MSA test (external variable). Groups, based on the mode of test administration, were then matched on the external variable (i.e., the scores on the MSA) in the analysis of mode effects on each item.

As stated earlier in this paper, most of the same students who took the Mod-MSA assessments in 2010 will have taken the P&P administered MSA in 2009. Since the MSA is a P&P only administration, scores of students on the MSA could be used as a strong unbiased variable for the creation of equivalent groups on the Mod-MSA for the DIF analysis. Because the Mod-MSA is similar to the MSA, student scores on the MSA can be seen as predictive of performance on the Mod-MSA, provided the items on the Mod-MSA indeed has no test-mode effects.

Since the Mod-MSA examinations do not have any polytomously scored items, the Mantel-Haenszel Chi-Square ($MH\chi^2$) together with ETS's Delta Scale were used for the contingency and the effect-size approach[1] to DIF.

The Mantel and Haenszel (1959) chi-square, which approximately follows a chi-square distribution with one degree of freedom, can be formulated as per the following (from Camilli & Shepard, 1994):

$$\text{MH } \chi^2 = \frac{\left\{\left|\sum_{j=1}^{S}\left[A_j - E\left(A_j\right)\right]\right| - 1/2\right\}^2}{\sum_{j=1}^{S} VAR\left(A_j\right)}, \text{ where}$$

A_j and $E(A_j)$ are the observed number of correct responses and the expected number on the item, respectively for the Reference group, while $VAR(A_j)$ is the variance associated with the observed score.

In order to calculate the Delta scale, the Mantel and Haenszel (1959) log odds ratio was calculated using the following equation:

$$\alpha_{MH} = \frac{\sum_{j=1}^{S} A_j D_j / T_j}{\sum_{j=1}^{S} B_j C_j / T_j}, \text{ where}$$

the various variables in the equation are from the following 2 x 2 contingency table for the *jth* total score on the test (Camilli & Shepard, 1994, p. 106).

[1] For a detailed discussion on Mantel-Haenszel Chi-square, the Delta Scale and ETS Categories, please refer to Camilli and Shepard (1994).

Score on studied item with general notation

Group		1	0	Total
	R	A_j	B_j	n_{Rj}
	F	C_j	D_j	nFj
		m_{1j}	m_{0j}	T_j

The log odds ratio is a transformation of the odds ratio with its range being in the interval $-\infty$ to $+\infty$. The simple natural logarithm transformation of this odds ratio is symmetrical around zero, in which zero has the interpretation of equal odds. The odds ratio is transformed into a log odds ratio as per the following: $\beta_{M-H} = \ln(\alpha_{M\text{-}H})$. β_{M-H}, also has the advantage of being transformed linearly to other interval scale metrics (Camilli & Shepard, 1994). This fact is utilized in creating the Delta scale (*D*), which is defined as $D = -2.35\beta_{M-H}$.

The *M-H* χ^2 is examined in conjunction with the Delta scale (*D*) to obtain DIF classifications depicted in Table 1, below.

Table 1: DIF Classification

Category	Description	Criterion
A	No DIF	Non-significant *M-H* χ^2 or $\|D\| < 1.0$
B	Weak DIF	Significant *M-H* χ^2 and $\|D\| < 1.5$ or Non-significant *M-H* χ^2 and $\|D\| > 1.0$
C	Strong DIF	Significant *M-H* χ^2 and $\|D\| \geq 1.5$

As stated previously, the groupings for the DIF analysis were based on matching students' scores on the MSA. Four proficiency-groupings of the Mod-MSA students were formed at quarter intervals of the total MSA score. All the students who had taken the Mod-MSA were used in this analysis. The Performance on the Mod-MSA for the four external proficiency-matched groups was then compared for each item to evaluate potential differential performance by mode.

The matching method described above for forming equal proficiency groups are the same as those used in conventional DIF analysis with one exception: instead of classifying proficiency groupings based on student performance on the test they have taken, i.e., the Mod-MSA, the four proficiency groupings were classified on the basis of their performance on the MSA. As explained earlier in this paper, this procedure allowed us to bypass the possible confounding effects on student abilities based on systemic differences between the modes of administration on the Mod-MSA tests, keeping in mind that the MSA on which the proficiency groups were classified is a P&P administered test only.

The DIF items identified by this procedure could then to be used to identify *biased* items (with respect to the testing modes) for *future* test forms development.

Results

As stated earlier, Grade 3 mathematics and reading were not included in the analyses because these students did not have the corresponding MSA scores from 2009. The main matching criterion was the student's State ID. Based on this criterion the samples for Grades 4 and 5 were adequate for our analysis and are depicted in Table 2, below. The respective mean and standard deviation on the Mod-MSA and the MSA for the students' performance are also displayed in the table. We used the SAS statistical program with the Proc GLM option to obtain the adjusted means and the corresponding F-values for the test of homogeneity of slopes, and significance of the main effect, i.e., the equality of the adjusted means between mode groups

Table 2: Descriptive Statistics by Grade and Content for the Mod-MSA Students who were Identified as Having a Corresponding Score on the MSA.

Subject	Grade	Type	N-Count	Mean Mod-MSA	Std. Dev. Mod-MSA	Mean MSA	Std. Dev. MSA
Mathematics	4	All	1184	26.97	8.31	31.77	11.56
		OL	268	26.12	7.27	32.56	10.53
		P&P	916	27.21	8.58	31.54	11.84
Mathematics	5	All	1290	25.12	7.36	25.98	9.66
		OL	325	23.71	6.16	25.10	8.82
		P&P	965	25.60	7.66	26.28	9.91
Reading	4	All	1225	24.83	7.04	11.69	5.09
		OL	276	24.61	6.42	12.90	5.43
		P&P	949	24.90	7.21	11.34	4.94
Reading	5	All	1337	24.74	6.64	16.89	5.85
		OL	335	24.27	6.14	16.87	5.72
		P&P	1002	24.91	6.79	16.90	5.89

Test-Level Analysis

In order to ascertain the viability of using ANCOVA as an analytical method we first tested the linear correlation between the covariate (the students' 2009 MSA scores) with the Mod-MSA scores. The results are presented in Tables 3.

Table 3: Correlation Between the 2009 MSA and the 2010 Mod-MSA scores

Subject	Grade	Examination Type	N-Count	Correlation Coefficient Between the 2010 Mod-MSA and the 2009 MSA
Mathematics	4	Mod-MSA	1184	0.47
		MSA	1184	-
Mathematics	5	Mod-MSA	1290	0.46
		MSA	1290	-
Reading	4	Mod-MSA	1225	0.36
		MSA	1225	-
Reading	5	Mod-MSA	1337	0.47
		MSA	1337	-

As can be seen from the above table, the correlations range from a low of 36 to a high of 47. Because of the restriction of range of the Mod-MSA student scores, the correlations may be lower than what would be expected if no attenuation had taken place.

The second consideration in the use of the ANCOVA, as discussed above, was the verification of the assumption of equality of the regression slopes (i.e. to test the testing-mode groups' interaction with the MSA scores). These results are presented in Table 4. As can be seen from the table, the homogeneity of the regression slopes is tenable across all grades and content at the 0.05 significance level.

Table 4: Assessing the Equality of the Regression Slopes

Subject	Grade	Source	DF	F-Value	Pr > F
Mathematics	4	MSA × Mode	1	0.85	0.3555
Mathematics	5	MSA × Mode	1	3.34	0.0679
Reading	4	MSA × Mode	1	0.57	0.4513
Reading	5	MSA × Mode	1	0.07	0.7985

Based on the homogeneity of the regression slopes, as indicated in the table above, we used the ANCOVA to test the difference between the adjusted means of the two mode-administered groups without having to resort to such techniques as the Johnson-Neyman method to establish the limits of the regions of non-significance on the covariate.

The adjusted means and the main effect significance table are provided in Tables 5 and 6, respectively. Table 6 also provides the magnitude of the difference between the adjusted means (i.e., the effect size (ES) measures).

Table 5: Adjusted Means of OL and P & P Groups

Subject	Grade	Adjusted Mean OL	Adjusted Mean P&P
Mathematics	4	25.85	27.30
Mathematics	5	24.02	25.50
Reading	4	23.99	25.08
Reading	5	24.28	24.90

Table 6: The *F*-Test for the Main Effects of ANCOVA: Testing for Equality of the Adjusted Means Between Mode Groups

Subject	Grade	N-Count	Source	DF	F-Value	Pr > F	Effect Size (ES) Measure
Mathematics	4	1184	Mode Groups	1	8.12	0.0045	0.08
Mathematics	5	1290	Mode Groups	1	12.57	0.0004	0.10
Reading	4	1225	Mode Groups	1	5.83	0.0159	0.07
Reading	5	1337	Mode Groups	1	2.84	0.0922	0.05

As can be seen from Table 5, the adjusted means for the P&P are higher than the OL for each of the content areas across grades (almost negligible for Grade 5 reading), indicating that on an average, groups that took the P&P performed better than those students who took the OL. Since the main effects are significant (Table 6), we rejected the null of no difference between mode-groups at the predetermined 0.05 level for all grades and content areas except for Grade 5 reading where there was no statistically significant difference between modes of administration.

However, in practical terms, the difference in the adjusted means is small as displayed by the effect-size (ES) measures shown in Table 6. The ES for the main effect was calculated by the following formula (Stevens, 1990, p. 143):

$ES = \sqrt{(k-1)F/N}$, where

k= level of the groups (which in our case = 2), and the F and N values are those that are shown in Table 6 above. The ES values depicted in the table are small as characterized by Cohen (1977) where and ES of around 0.10 is considered small, around 0.25 as medium, and .0.40 as large (Stevens, 1990, p. 89).

Item-Level Analysis

Prior to analyzing items for DIF, simple p-value charts (Figure 1 to Figure 4) that reflect each item's performance between the modes of administration by grades and content is provided below. These charts give a general idea on item behavior across modes, keeping in mind that no adjustment was made with respect to the proficiency groupings of students between the two modes, and students were not assigned randomly to the modes of administration.

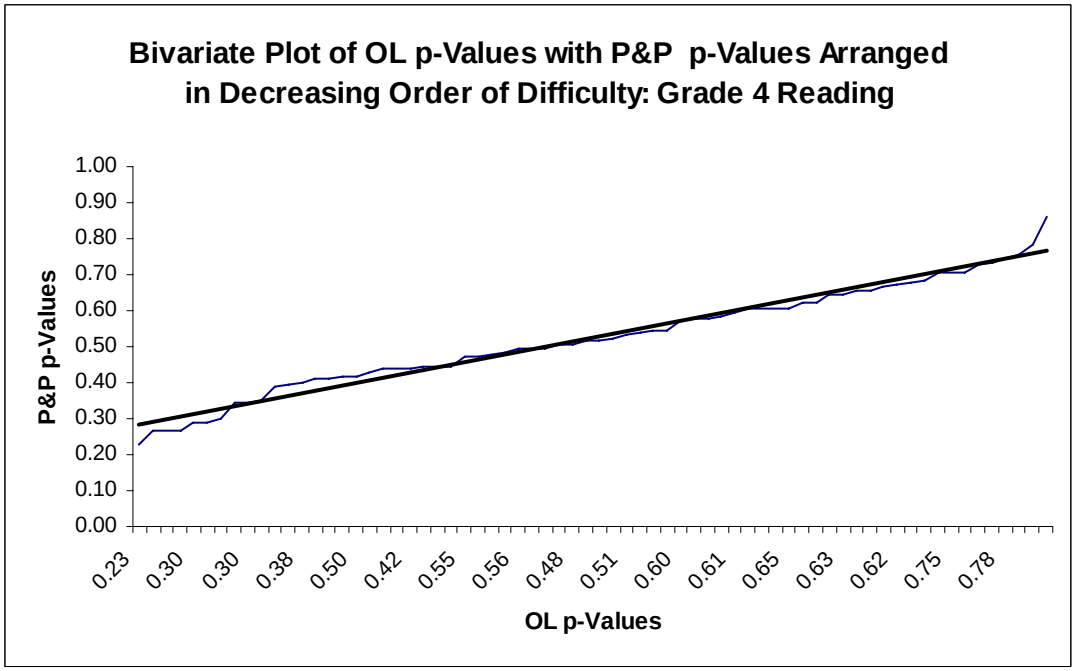

Figure 1: Grade 4 reading item p-values by mode of administration

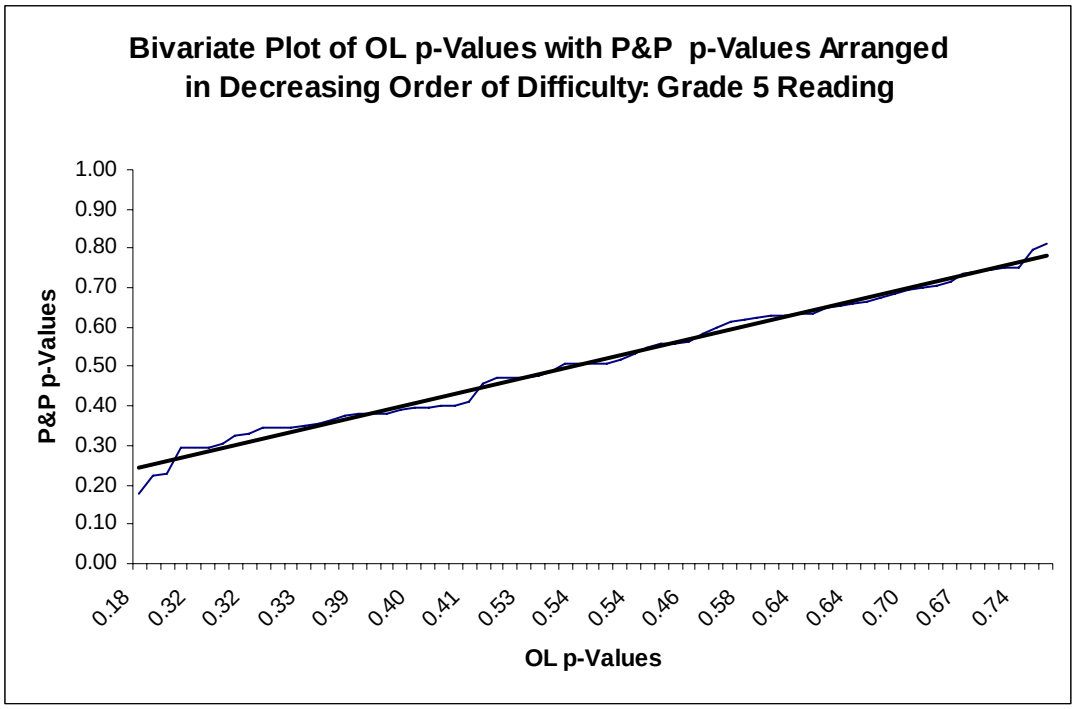

Figure 2: Grade 5 reading item p-values by mode of administration

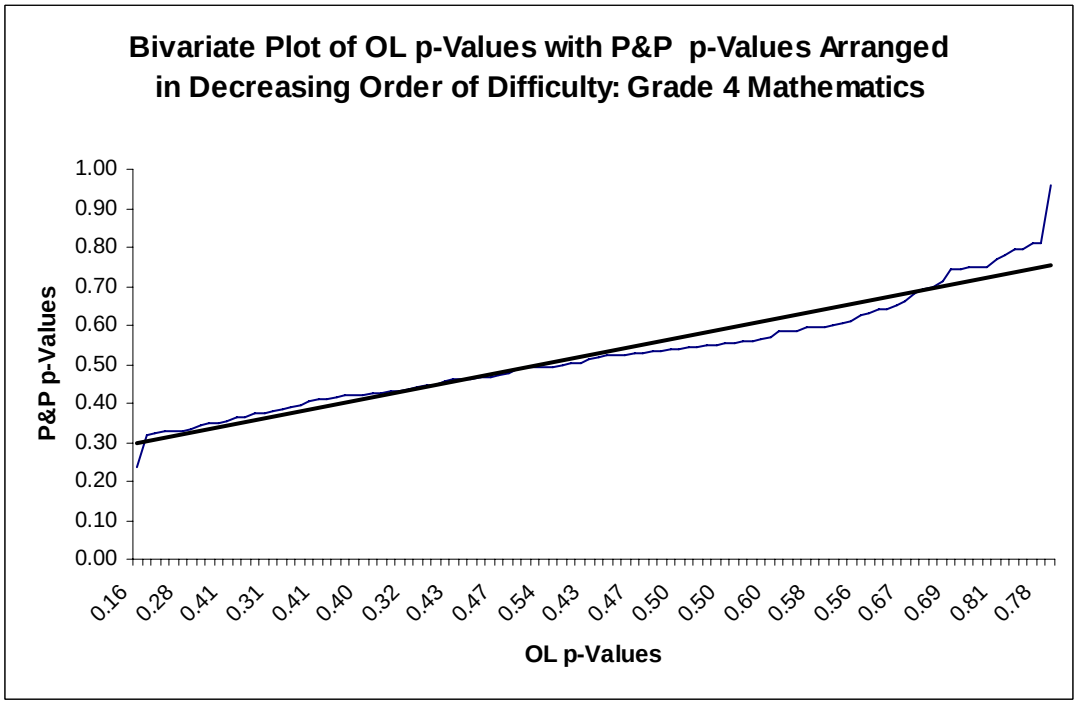

Figure 3: Grade 4 mathematics item p-values by mode of administration

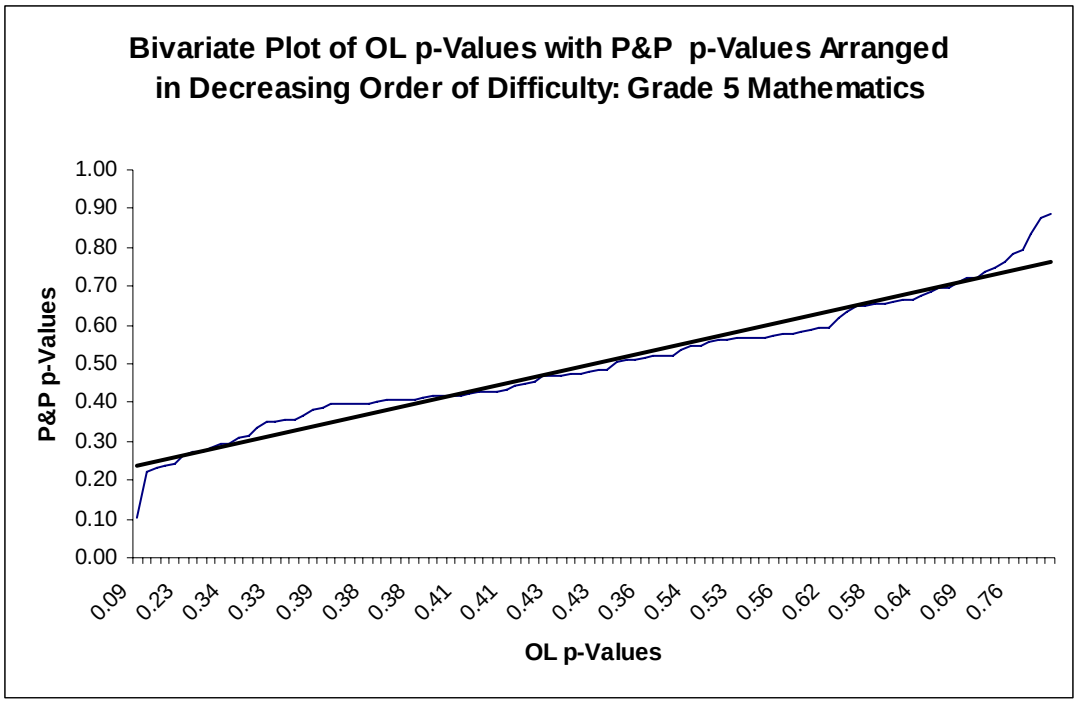

Figure 4: Grade 5 mathematics item p-values by mode of administration

By using the 2009 MSA scores as an external variable for matching the mode-administered groups, we found no extreme category, "C" DIF for any of the items (see Table 7, below), which included both operational items and those items that were not used as core items. For this analysis, the SAS program was once again used to calculate the M-H chi-square significance and the effect type measure of the delta scale.

Table 7: DIF Classification of Flagged Items by Content and Grade

Subject	Grade	Item Sequence No.	Item CID No.	M-H Chi-Square	Chi-Square Probability	Delta Scale	DIF Category
Mathematics	4	67	100000186573	9.85	0.002	1.10	-B
Mathematics	4	84	100000198121	11.68	0.001	1.17	-B
Mathematics	5	73	100000196094	21.13	0.00	1.42	-B
Reading	4	17	100000357136	10.76	0.001	1.07	-B
Reading	5	12	100000102095	12.38	0.000	1.05	-B

Note: + = in favor of P&P and - = in favor of OL.

All the items that were administered and scored for Grades 4 and 5 reading and mathematics were used in the DIF analysis (i.e., a total of 337 items broken down by 102 items in Grade 4 and 100 items in Grade 5 mathematics, and 68 and 67 items for reading in Grades 4 and 5 respectively). As shown in Table 7, there were two category "B" DIF classifications for mathematics for Grade 4. The remaining grades in reading had one each with "B" classification. All the items with DIF were in favor of OL.

Comments and Conclusion

The methods described in this study can be seen as two approaches to test the same null hypothesis of no examinee differences in student performance between test modes. However, the results of the two methods have different implications in assessing the impact of testing modes on students who are administered the Mod-MSA. The ANCOVA provides an overall view of test-mode effects by considering the differences between the test-mode groups in terms of the *total* test performance. As such, the results from the analysis can be seen as the total of item effects.

The DIF analysis, on the other hand, tests the hypothesis of no difference between testing-mode groups at the item level. In a sense, the two approaches complement each other by the analysis of individual item behavior as in DIF and the total item behavior as in the ANCOVA.

In our analysis, statistically significant differences were found at three out of four grade levels within a content area for the ANCOVA main effect. It was also found that the differences on average were in favor of those who took the P&P (i.e. the test as a whole with the exception of Grade 5 reading, was slightly harder for OL test takers in comparison to those who took the P&P

testing mode). The significant differences, however, could be attributed partly to the attenuation effects for the Mod-MSA students. Greater precision of estimate would have been possible if the correlation between the independent and the dependent variable was not underestimated.

As we had discussed earlier, the assignment of students to the mode of administration was not random. In such cases, the ANCOVA (as in most other statistical analyses) has an important limitation (Anderson, 1963; Lord, 1969) that needs to be addressed. As Stevens (1990, p.168) points out: "even the use of several covariates will not equate intake groups, and one should not be deluded into thinking that it can. The groups may still differ on some unknown important variable(s)."

In this study, it is quite possible that a non-modeled variable(s) could have had an impact on the groups in question. For example, it is likely that school and student variables (e.g., the degree of schools' encouragement in the use of technology, student non-familiarity with computer testing, etc.) *may* have had an effect on student achievement. Future studies *modeling* these variables may provide some explanation of these hypothetical concerns.

However, it is important to avoid placing too much emphasis on these statistically significant results, as the actual differences between the adjusted means were small as measured by their effect sizes (ES). The variability of the adjusted group means about the grand mean as shown in Table 6 is small. Because of the large sample sizes with respect to the ANCOVA, the least amount of practically insignificant difference (e.g., differences so small as to have a negligible affect on student scores) between groups can show up as being statistically significant. It, therefore, makes sense to examine ES measures as a pragmatic approach in the comparison of mode effects for the Mod-MSA.

It is encouraging to note that the results of this study indicate that all the items used for the 2010 administration for Grades 4 and 5 across the two content areas did not show extreme DIF between modes of administration. The *moderate* DIF shown for a total of a mere five items from 337 items across grades and content areas (far fewer than would be expected by chance at alpha = 0.05) can be scrutinized for mode bias by content specialists.

The very small effect sizes and the relative absence of DIF suggest the viability of using P&P as a replacement for on-line administrations when needed. However, the MSDE is encouraged to continue mode DIF analysis for new items in the future to the extent that the availability of data makes such analysis possible.

References

Anderson, N. H. (1963). Comparison of different populations: Resistance to extinction and transfer. *Psychological Bulletin*, *70*, 162-179.

Braun, H., Jenkins, F., & Grigg, W. (2006). *Comparing private schools and public schools using hierarchical linear modeling*. (NCES 2006-461). Washington D.C: U. S. Department of Education.

Camilli, G., & Shepard, L. (1994). *MMS: methods for identifying biased test items.* Sage.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Eignor, D. R. (2007). Linking scores derived under different modes of test administration. In Dorans, N. J. , Pommerich, M., & Holland, P. W. (Eds.) *Linking and aligning scores and scales* (pp. 135-137). Springer, N.Y.

Garson, G. David (n.d.). Linear Mixed Models: Random effects, hierarchical linear, multilevel random coefficients, and repeated measure models, from *Statnotes: Topics in multivariate analysis.* Retrieved 06/26/2008 from http://www2.chass.ncsu.edu/garson/pa765/statnote.htm.

Glasnapp, D. R., Poggio, J., Poggio, A., & Yang, X. (2005). *Student attitudes and perceptions regarding computerized testing and the relationship to performance in large scale assessment programs*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Thousand Oaks, CA: Sage.

Hoffmann, D. A. (1991). An overview of the logic and rationale of hierarchical linear Models. *Journal of Management, 23*(6), 723-744.

Huang, M., & Lu, E. (2007). *The two-level sample size problem of hierarchical linear modeling: Evidence from simulation experiments*. Paper presented at the annual meeting of the American Political Science Association, Chicago, IL.

Kim, D. H., & Huynh, H. (2008). Computer-based and paper-and-pencil administration mode effects on a statewide end-of-course English test. *Educational and Psychological Measurement, 68(4)*, 554-570.

Kim, D. H., & Huynh, H. (2009). Transition from paper-and-pencil to computer-based testings: examining stability of Rasch latent trait across gender and ethnicity. To appear as a book chapter in Everett V. Smith Jr. & Greg E. Stone (Eds.) *Applications of Rasch measurement in criterion-referenced testing: Practice analysis to score reporting*. Maple Grove, MN: JAM Press.

Keng, L., McClarity, K. L., & Davis, L. L. (2006). *Item-level comparability analysis of on-line and paper administrations of the Texas Assessment of Knowledge and Skills*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating: methods and practices.* New York, NY: Springer-Verlag.

Linacre, J. M. (2001). *WINSTEPS Rasch measurement program, Version 3.32.* Chicago: John M. Linacre.

Lord, F. M. (1969). Statistical adjustments when comparing pre-existing groups. *Psychological Bulletin*, *70*, 162-179.

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

Pedhazur, E. J., (1973). *Multiple regression in behavioral research*. Forth Worth, TX: Harcourt Brace Jovanovich.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of scored results from computerized and Paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning and Assessment*, *3*(6).

Raudenbush, S. W. (1988) Educational applications of hierarchical Models: a review. *Journal of Educational Statistics, 13*, 85-116.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (pp. 238-345). Hillsdale, NJ: Lawrence Erlbaum.

U.S. Department of Education. Fed. Reg. 34 CFR Parts 200 and 300 (April 9, 2007). *Title I-Improving the academic achievement of the disadvantaged; Individual of Disabilities Education Act, Final rule*.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). *Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.