MISSOURI END-OF-COURSE
PAPER/PENCIL VERSUS ONLINE COMPARABILITY STUDY

Introduction

The Missouri End-of-Course (EOC) Assessments for English II, Algebra I, and Biology were developed by Riverside Publishing and first field tested in the spring of 2008. The first operational testing events occurred during the 2008/2009 school year. The EOC Assessments were created to be more targeted assessments to meet the needs of Missouri districts, schools, teachers, and students, while also meeting state and federal requirements.

Each MO EOC Assessment includes two types of test items: selected-response items and performance events (PE) or a writing task. The EOC Assessments are administered across two testing sessions of approximately one class period each. The selected-response items are administered in Session I, and the PE items are administered in Session II.

For each administration of the MO EOC Assessments, schools are given the option to administer the test in a paper-and-pencil (P/P) version, or to administer the test online. Approximately 10% of Missouri students took the EOC Assessments through the online administration option during the 2008–2009 school year.

The comparability of a computer-based assessment to its paper-and-pencil counterpart cannot be assumed. Conceivably, the mode of administration may affect the difficulty of the test, either through an overall shift in difficulty or through an item-by-mode interaction. Riverside Publishing conducted the current study for the purpose of describing a strategy for evaluating the comparability of Missouri's P/P and online EOC Assessments and to provide a summary of several analyses performed to determine the comparability of the two modes for the Spring 2009 administration. Because a relatively small number of students took the online assessment in Fall 2008, a comparability study could not be performed for that test administration.

Because of the potential confounding caused by sampling bias, and to help interpret the results in this context, this study employed two different types of samples and a number of different analyses. These analyses are presented as a "body of evidence" to assist in evaluating the potential effect of mode of administration on test results.

Sample

A specific challenge for the evaluation of comparability between the two modes of administration for the MO EOC Assessments was that the samples of students taking the test in each mode were not randomly equivalent. Participation in the online administration was voluntary; thus, the only students who took the test online were those from schools or districts that self-selected for online administration. In short, students who participated in the online testing were not representative of the total population of students in Missouri.

Any analyses using the entire data set ("total sample") would be impacted by this nonequivalence. More specifically, results from analyses on the total sample could be confounded by differences that might exist in the two samples due to sampling bias.

Copyright © by Missouri Department of Elementary and Secondary Education

Nevertheless, comparisons based on the total sample (all paper/pencil versus online test-takers) were performed to set a baseline for the differences between the online and paper/pencil assessments.

In an attempt to control for the differences in student ability and other demographic characteristics between the two samples, a second approach for data analysis was taken. In this approach, a "matched sample" was created from the larger total sample. In this matched sample, each student who took the test online was matched on important demographic variables with a student who took the P/P version of the test, so that the result was a sample of test-takers with more equivalent characteristics than the original total sample. The variables used to match the students from the P/P and online samples were content-area MAP scale scores, student grade level, and participation in free and reduced-price lunch programs (FRL).

The following steps were used to match students in each content area:

1. Using a student-level database containing grade-level content-area MAP scale scores, isolate the most recently administered MAP grade-level assessment for each student.
2. Create a student "matching variable" that is the concatenation of student grade level, MAP scale score, and participation in FRL.
3. Sort all online students and all paper/pencil students by the matching variable. Combine the two datasets by merging with the matching variable.
4. Create a uniform random variable and sort by student ID and the random variable. (Because the sample of paper/pencil test-takers was much larger than the sample of online test-takers, each online student matched with multiple P/P students. Therefore, this step was necessary to randomly select one paper/pencil student for each online student.)
5. Select the first paper/pencil student matched with each online student.

The percentage of matched cases for each MO EOC Assessment is included in Table 1. Note that the percentage of matched students who took the online Biology assessment was significantly lower than that of the English II and Algebra I assessments. This is because of the limited number of MAP grade-level Science scores available for the total sample. Because the MAP grade-level Science assessment was not required until the 2007-2008 school year, only students taking the Biology EOC in Grade 9 in Spring of 2009 had grade-level MAP Science scores from Grade 8 in the data file. Overall, however, the limited number of variables used for matching (three) and the high (ten to one) ratio of P/P to online test-takers led to a high overall percentage of matches when MAP scores were available.

Copyright © by Missouri Department of Elementary and Secondary Education

Table 1. Percentage of Matched Cases in the Online and P/P Matched Samples for Each EOC Assessment

	Total number of Students in Online Sample	Number of Online Students Matched with a Paper/Pencil Student	Percentage of Online Students Matched
English II	6,837	5,832	85%
Algebra I	3,956	3,678	93%
Biology	6,343	1,462	23%

Tables 2 through 5 provide distributions for ethnicity for the total sample and the matched sample. Distributions by EOC Assessment (i.e., English II, Algebra I, and Biology) had similar percentages and thus are not reported individually. Table 6 provides the grade-level distribution for each matched sample of paper/pencil and online test takers, and Table 7 provides the distribution of free and reduced lunch status for each matched sample. Table 8 provides the mean MAP scale score for each matched sample. Recall that the variables reported in Tables 6 through 8 were used for matching students. Because each student in the online sample corresponded perfectly to a student in the P/P sample on the three matching variables, the values reported in Tables 6 through 8 apply to both the matched online sample and the matched P/P sample.

Table 2: Ethnicity Distribution for Total Sample of Paper/Pencil Test Takers

Ethnicity	*N*	Percent
African American	20,523	13.9
Asian	714	0.4
Asian/Pacific Islander	2,737	1.8
Hispanic	4,261	2.8
White	119,222	80.8
Total	147,457	

Note: Percentages may not sum to 100 due to rounding.

Table 3: Ethnicity Distribution for Total Sample of Online Test Takers

Ethnicity	*N*	Percent
African American	4,276	23.2
Asian	71	0.4
Asian/Pacific Islander	265	1.4
Hispanic	555	3.0
White	12,227	71.9
Total	17,394	

Note: Percentages may not sum to 100 due to rounding.

Table 4: Ethnicity Distribution for the Matched Sample of Paper/Pencil Test Takers

Ethnicity	*N*	Percent
African American	1,695	15.5
Asian	39	0.4
Asian/Pacific Islander	202	1.8
Hispanic	313	2.9
White	8,723	79.5
Total	10,972	

Note: Percentages may not sum to 100 due to rounding.

Copyright © by Missouri Department of Elementary and Secondary Education

Table 5: Ethnicity Distribution for the Matched Sample of Online Test Takers

Ethnicity	*N*	Percent
African American	2,256	20.6
Asian	41	.40
Asian/Pacific Islander	128	1.2
Hispanic	328	3.0
White	8,219	74.9
Total	10,972	

Note: Percentages may not sum to 100 due to rounding.

Table 6: Grade-Level Distribution for each Matched Sample of Paper/Pencil and Online Test Takers

Grade Level	*N*	Percent
7	7	0.1
8	698	6.4
9	3,749	34.2
10	6,204	56.5
11	142	1.3
12	172	1.6

Note: Percentages may not sum to 100 due to rounding.

Table 7: Distribution of Free and Reduced Lunch Status for each Matched Sample of Paper/Pencil and Online Test Takers

Grade Level	*N*	Percent
Free and Reduced Lunch	4,076	37.2
Not Free and Reduced Lunch	6,896	62.9

Note: Percentages do not sum to 100 due to rounding.

Table 8: Mean MAP Scale Score for each Matched Sample of Paper/Pencil and Online Test Takers

Grade Level	*N*	Mean	SD
English II	5,832	694.5	32.3
Algebra I	3,678	707.3	33.7
Biology	1,462	703.2	26.1

To verify the representativeness of the matched samples with respect to other possible matched samples obtained from the population of online and paper/pencil test takers, the original matching process was replicated an additional 19 times. Results summarizing the means from all 20 matched samples (i.e., replications) generally support the use of the *first* or original matched sample that was selected and used for subsequent analyses.

Methods and Results

This section provides a summary of several analyses completed to investigate whether the online and paper/pencil versions of the MO EOC Assessments administered in Spring 2009 were comparable. Because the samples of students who took the test via each mode of administration were not randomly equivalent, a clear "yes" or "no" answer to the question of comparability was not possible. Rather, various analyses were employed to lend evidence to either support or refute the hypothesis that the test scores obtained from each mode of administration were comparable.

Copyright © by Missouri Department of Elementary and Secondary Education

Table 9 presents an overview of the analyses included in this study, separated by two different approaches to using the available sample data.

Table 9: Analyses Used to Evaluate the Comparability of Paper/Pencil and Online Administrations

Sample Approach		Purpose
Total Sample	**Matched Sample**	
Comparison of Summary Statistics and Item Means	Comparison of Summary Statistics and Item Means	Evaluate the observed differences in student performance between the P/P and online groups
ANCOVA		Evaluate the differences in mean student performance after statistically removing the effects of the covariates
	Comparison of Item Difficulties	Examine individual item difficulties within each mode
Differential Item Functioning Analysis	Differential Item Functioning Analysis	Using the online students as the focal group and the P/P students as the reference group, identify and review items that appear to function differently for the two modes of administration
	Hypothetical Equating	Evaluate the practical magnitude of potential mode effects by examining raw score-to-scale score tables as if the online administration was treated as a unique form equated to the P/P administration
	Confirmatory Factor Analysis	Evaluate the comparability of factor structures for the P/P and online administrations of the tests

The following sections describe the purpose of each analysis and detail the results and implications of each.

Comparison of Summary Statistics for the Total Sample and Matched Sample

To compare the total sample and the matched sample, summary statistics, including mean score, standard deviation, and minimum/maximum scores were calculated for each mode of administration in the total sample and the matched sample for each assessment. Tables 10 and 11 present summary statistics for the two samples. Note that mean raw score differences between paper/pencil and online modes are larger in Table 10, which contains the data for the total sample (all P/P and online test-takers). Recall that in this group, the online sample of students is not representative of the total student population. Differences generally become smaller (less than one raw score point) for the matched sample. These smaller differences in overall test scores between the two modes in the matched sample suggest that the lack of representation in the total sample did contribute to differences between the two modes.

In addition to the summary statistics for the first matched sample created from the total data set, Table 11 also provides mean summary statistics for all 20 replications. Note that

Copyright © by Missouri Department of Elementary and Secondary Education

the first matched sample created is representative of the mean of the replications. Thus, further analysis was done using only the first matched sample created.

Table 10: Summary Statistics for the Total Sample of Paper/Pencil and Online Test Takers

Content Area/ Mode of Administration	*N*	Mean	*SD*	Minimum	Maximum
English II					
Paper/Pencil	49,843	27.5	6.2	3	39
Online	6,837	26.3	6.1	5	39
Algebra I					
Paper/Pencil	48,622	21.8	7.0	2	38
Online	3,956	20.4	6.9	4	38
Biology					
Paper/Pencil	48,992	33.1	9.7	3	55
Online	6,343	32.2	9.6	5	55

Table 11: Summary Statistics for the Matched Sample of Paper/Pencil and Online Test Takers

Content Area/ Mode of Administration	***N***	**Mean**	***SD***	**Minimum**	**Maximum**
First Matched Sample					
English II					
Paper/Pencil	5,832	27.4	6.1	6	39
Online	5,832	26.6	6.0	5	39
Algebra I					
Paper/Pencil	3,678	21.4	7.1	2	38
Online	3,678	20.5	6.9	4	38
Biology					
Paper/Pencil	1,462	32.6	9.8	6	52
Online	1,462	32.2	9.7	6	52
All 20 Matched Samples					
English II					
Paper/Pencil	20	27.4	.049	27.3	27.5
Online	20	26.6	.001	26.6	26.6
Algebra I					
Paper/Pencil	20	21.4	.055	21.3	21.5
Online	20	20.5	.000	20.5	20.5
Biology					
Paper/Pencil	20	32.5	.136	32.3	32.8
Online	20	32.2	.006	32.2	32.2

To further support the results obtained from the matched sample, an analysis of covariance (ANCOVA) was performed to statistically remove variability associated with the three matching variables (i.e., MAP scale score, grade level, and FRL). More specifically, ANCOVA was used to provide verification of the matching results. The means adjusted for the covariates were obtained for comparison—and verification–of the means obtained with the matched samples. Additionally, ANCOVA can uncover the relative statistical contribution of each of the matching variables. Tables 12 through 14 present the results of the ANCOVA for each EOC Assessment. Table 15 provides the total raw score means adjusted for the covariates.

Copyright © by Missouri Department of Elementary and Secondary Education

Table 12: ANCOVA Results for English II

Source	SS	df	MS	*F*	Sig.
Corrected Model	248025.4	4	62006.4	3766.4	<.001
Error	193226.2	11737	16.5		
Corrected Total	441251.6	11741			
Mode (Online/Paper)	2270.4	1	2270.4	137.9	<.001
Map Scale Score	213108.7	1	213108.7	12944.7	<.001
Grade	217.1	1	217.1	13.2	<.001
FRL	2688.1	1	2688.1	163.3	<.001
R^2 = .56					

Table 13: ANCOVA Results for Algebra I

Source	SS	df	MS	*F*	Sig.
Corrected Model	214463.1	4	53615.8	2776.1	<.001
Error	143325.6	7421	19.3		
Corrected Total	357788.7	7425			
Mode (Online/Paper)	1268.3	1	1268.3	65.7	<.001
Map Scale Score	140297.8	1	140297.8	7264.2	<.001
Grade	3202.0	1	3202.0	165.8	<.001
FRL	1116.3	1	1116.3	57.8	<.001
R^2. = .60					

Table 14: ANCOVA Results for Biology

Source	SS	df	MS	*F*	Sig.
Corrected Model	184698.4	4	46174.6	1401.7	<.001
Error	98660.3	2995	32.9		
Corrected Total	283358.6	2999			
Mode (Online/Paper)	61.5	1	61.5	1.87	<.172
Map Scale Score	150227.5	1	150227.5	4560.4	<.001
Grade	4355.1	1	4355.1	132.2	<.001
FRL	2641.7	1	2641.7	80.19	<.001
R^2 = .65					

Table 15: ANCOVA Adjusted Means

Content Area/ Mode of Administration	Adjusted Means	Standard Error
English II		
Paper/Pencil	27.5	.053
Online	26.6	.053
Algebra I		
Paper/Pencil	21.5	.072
Online	20.7	.072
Biology		
Paper/Pencil	33.3	.149
Online	33.0	.149

Copyright © by Missouri Department of Elementary and Secondary Education

To balance the ANCOVA analysis, a random sample of P/P examinees equal to the number of online examinees was selected. An examination of Tables 12 through 14 shows that MAP scale scores are by far the most important covariate in the analysis and also in the matching process described earlier. Perhaps more importantly, the mean scores adjusted for the covariates are very similar in both magnitude and their respective differences when compared to the mean scores from the matched samples. Both analyses indicate that when the effects of the covariates or matching variables are statistically removed or balanced, there is a notable reduction in the differences between EOC Assessments' score means.

Comparison of Item Difficulties for the Matched Sample

One way to determine whether two modes of administration are equivalent is to examine individual item difficulties within each mode. Using the matched sample, item difficulties in the form of *p*-values were computed for each item within each mode of administration. For PE items, mean item scores were computed. Similar item difficulties between corresponding online and P/P items in the matched sample would suggest that the test item difficulties were not affected by mode of administration.

Tables 16 through 21 present comparisons between the item *p*-values and item means (for the PE items) for the matched samples. For each content area, the difference between the P/P and online item *p*-values is also provided (Tables 16, 18 and 20). In addition, the frequency of differences between *p*-values for each matched sample is given (Tables 17, 19 and 21). Differences between *p*-values were generally small, falling within the range of –.05 to .05. A few items did show larger differences, perhaps indicating an interaction between item difficulty and mode of administration.

Table 16: Difference between *p*-Values/Item Means for the Matched Sample–English II

Item	*p*-Value for Paper/Pencil	*p*-Value for Online	Paper/Pencil Minus Online *p*-Value
Item 1	0.74	0.69	0.05
Item 2	0.74	0.74	0.00
Item 3	0.88	0.86	0.02
Item 4	0.89	0.80	0.09
Item 5	0.90	0.88	0.02
Item 6	0.37	0.33	0.04
Item 7	0.85	0.85	0.00
Item 8	0.74	0.70	0.04
Item 9	0.48	0.45	0.03
Item 10	0.75	0.70	0.05
Item 11	0.84	0.80	0.04
Item 12	0.53	0.56	-0.03
Item 24	0.58	0.60	-0.02
Item 25	0.56	0.53	0.03
Item 26	0.86	0.86	0.00
Item 27	0.55	0.47	0.08

Copyright © by Missouri Department of Elementary and Secondary Education

Table 16: Difference Between *p*-Values/Item Means for the Matched Sample–English II (continued)

Item 28	0.66	0.67	-0.01
Item 29	0.46	0.48	-0.02
Item 30	0.44	0.39	0.05
Item 31	0.81	0.78	0.03
Item 32	0.75	0.70	0.05
Item 33	0.45	0.37	0.08
Item 34	0.64	0.63	0.01
Item 35	0.89	0.86	0.03
Item 36	0.92	0.90	0.02
Item 37	0.61	0.58	0.03
Item 38	0.58	0.54	0.04
Item 39	0.85	0.82	0.03
Item 40	0.64	0.62	0.02
Item 41	0.71	0.73	-0.02
Item 43	0.87	0.86	0.01
Item 44	0.54	0.52	0.02
Item 45	0.81	0.81	0.00
Item 46	0.79	0.80	-0.01
Item 47	0.74	0.72	0.02
Item 48*	3.03	2.92	0.11

*Item 48 is a performance event.

Table 17: Frequency of Differences between *p*-Values for the Matched Sample–English II

Difference (Paper/Pencil Minus Online *p*-Value)	Frequency	Percent
-0.03	1	2.8
-0.02	3	8.3
-0.01	2	5.6
0.00	4	11.1
0.01	2	5.6
0.02	6	16.7
0.03	6	16.7
0.04	4	11.1
0.05	4	11.1
0.08	2	5.6
0.09	1	2.8
0.11	1	2.8

Copyright © by Missouri Department of Elementary and Secondary Education

Table 18: Difference between *p*-Values and Item Means for the Matched Sample–Algebra I

Item	*p*-Value for Paper/Pencil	*p*-Value for Online	Paper/Pencil Minus Online *p*-Value
Item 1	0.9	0.9	0.00
Item 2	0.79	0.76	0.03
Item 3	0.72	0.69	0.03
Item 4	0.73	0.71	0.02
Item 5	0.72	0.68	0.04
Item 10	0.82	0.78	0.04
Item 11	0.70	0.68	0.02
Item 12	0.74	0.71	0.03
Item 13	0.60	0.57	0.03
Item 14	0.43	0.42	0.01
Item 15	0.67	0.66	0.01
Item 16	0.65	0.65	0.00
Item 17	0.81	0.84	-0.03
Item 18	0.81	0.77	0.04
Item 19	0.65	0.63	0.02
Item 20	0.55	0.57	-0.02
Item 21	0.51	0.53	-0.02
Item 27	0.49	0.40	0.09
Item 28	0.55	0.52	0.03
Item 29	0.52	0.50	0.02
Item 30	0.52	0.50	0.02
Item 31	0.57	0.57	0.00
Item 32	0.48	0.45	0.03
Item 33	0.40	0.31	0.09
Item 34	0.48	0.47	0.01
Item 35	0.61	0.59	0.02
Item 36	0.37	0.32	0.05
Item 37	0.57	0.54	0.03
Item 38	0.37	0.33	0.04
Item 43	0.31	0.29	0.02
Item 44	0.58	0.58	0.00
Item 45	0.41	0.38	0.03
Item 46	0.12	0.08	0.04
Item 47	0.33	0.27	0.06
Item 48*	1.92	1.84	0.08

*Item 48 is a performance event.

Copyright © by Missouri Department of Elementary and Secondary Education

Table 19: Frequency of Differences between *p*-Values for the Matched Sample–Algebra I

Difference (Paper/Pencil Minus Online *p*-Value)	Frequency	Percent
-0.03	1	2.9
-0.02	2	5.7
0	4	11.4
0.01	3	8.6
0.02	7	20.0
0.03	8	22.9
0.04	5	14.3
0.05	1	2.9
0.06	1	2.9
0.08	1	2.9
0.09	2	5.7

Table 20: Difference between *p*-Values/Item Means for the Matched Sample–Biology

Item	*p*-Value for Paper/Pencil	*p*-Value for Online	Paper/Pencil Minus Online *p*-Value
Item 1	0.81	0.74	0.07
Item 2	0.79	0.76	0.03
Item 3	0.91	0.91	0.00
Item 4	0.66	0.69	-0.03
Item 5	0.83	0.8	0.03
Item 10	0.64	0.64	0.00
Item 11	0.72	0.71	0.01
Item 12	0.57	0.56	0.01
Item 13	0.66	0.61	0.05
Item 14	0.46	0.47	-0.01
Item 15	0.58	0.66	-0.08
Item 16	0.76	0.72	0.04
Item 17	0.93	0.92	0.01
Item 18	0.68	0.65	0.03
Item 19	0.48	0.48	0.00
Item 20	0.33	0.35	-0.02
Item 21	0.58	0.54	0.04
Item 26	0.73	0.72	0.01
Item 27	0.68	0.59	0.09
Item 28	0.50	0.43	0.07
Item 29	0.45	0.42	0.03
Item 30	0.41	0.35	0.06
Item 31	0.44	0.38	0.06
Item 32	0.59	0.57	0.02
Item 33	0.62	0.58	0.04
Item 34	0.58	0.56	0.02
Item 35	0.72	0.66	0.06
Item 36	0.62	0.58	0.04
Item 37	0.50	0.54	-0.04
Item 38	0.71	0.67	0.04

Copyright © by Missouri Department of Elementary and Secondary Education

Table 20: Difference Between *p*-Values/Item Means for the Matched Sample–Biology (continued)

Item 43	0.61	0.58	0.03
Item 44	0.55	0.55	0.00
Item 45	0.41	0.42	-0.01
Item 46	0.58	0.52	0.06
Item 47	0.84	0.83	0.01
Item 48*	0.66	0.66	0.00
Item 49*	0.79	0.84	-0.05
Item 50*	0.73	0.79	-0.06
Item 51*	2.53	2.79	-0.26
Item 52*	0.75	0.72	0.03
Item 53*	0.94	1.02	-0.08
Item 54*	1.14	1.08	0.06
Item 55*	0.74	0.77	-0.03
Item 56*	0.66	0.67	-0.01
Item 57*	0.97	1.08	-0.11
Item 58*	0.84	0.71	0.13

*Items 48 through 58 are performance event items.

Table 21: Frequency of Differences between *p*-Values for the Matched Sample–Biology

Difference (Paper/Pencil Minus Online *p*-Value)	Frequency	Percent
-0.26	1	2.2
-0.11	1	2.2
-0.08	2	4.4
-0.06	1	2.2
-0.05	1	2.2
-0.04	1	2.2
-0.03	2	4.4
-0.02	1	2.2
-0.01	3	6.5
0.00	5	10.9
0.01	5	10.9
0.02	2	4.4
0.03	6	13.0
0.04	5	10.9
0.05	1	2.2
0.06	5	10.9
0.07	2	4.4
0.09	1	2.2
0.13	1	2.2

Comparison of the Raw Score-to-Scale Score Conversion Tables Based on the Total Sample

One practical way to evaluate whether the two modes are equivalent is to examine the raw score-to-scale score conversions obtained through the use of traditional equating methods. If the raw scores corresponding to each scale score were the same between the two modes of administration, it would suggest that students would not be disadvantaged by either mode of administration. Before the conversion tables could be created,

Copyright © by Missouri Department of Elementary and Secondary Education

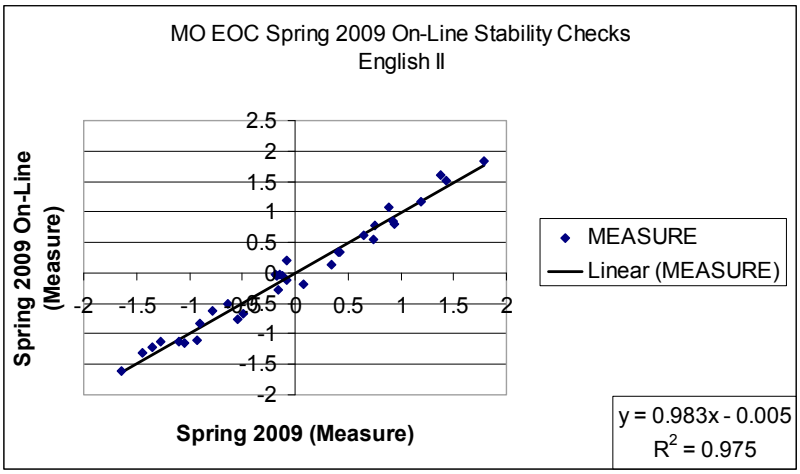
however, the two "forms" were brought onto the same scale using traditional equating methods. In a typical equating study, only a percentage of the items would be common to both forms. In this application, however, all the items are common between the online and P/P versions. In other words, this procedure was similar to a common-item equating with the potential for all items to be in common.

In this application, the P/P was considered the "base" form. To bring the online administration onto the P/P scale, the online items were anchored to their Spring 2009 operational item difficulty values, and the online data were recalibrated. Displacement values were examined for each item, and items with displacements greater than 0.3 logits were removed from the "common" set of items and allowed to calibrate freely. The displacement values were examined again, and the process repeated until a stable set of common anchor items was established. Table 22 shows the number of items dropped from the set of operationally administered items (i.e., the common set of items) for each assessment due to instability. Figures 1 through 3 show scatterplots of the final set of common items used to bring the online "form" onto the paper/pencil scale.

Table 22: Number of Unstable Items Dropped from the Common Set of Anchor Items for Each Assessment

Assessment	Number of Items Dropped
English	2
Algebra	1
Biology	1

Figure 1. Scatterplot of the Common Items Between Online and P/P Administrations: English II



Copyright © by Missouri Department of Elementary and Secondary Education

Figure 2. Scatterplot of the Common Items Between Online and P/P Administrations: Algebra I

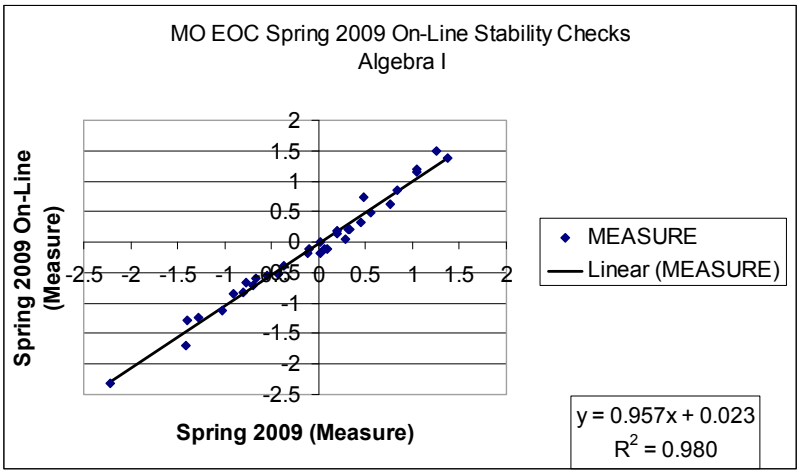
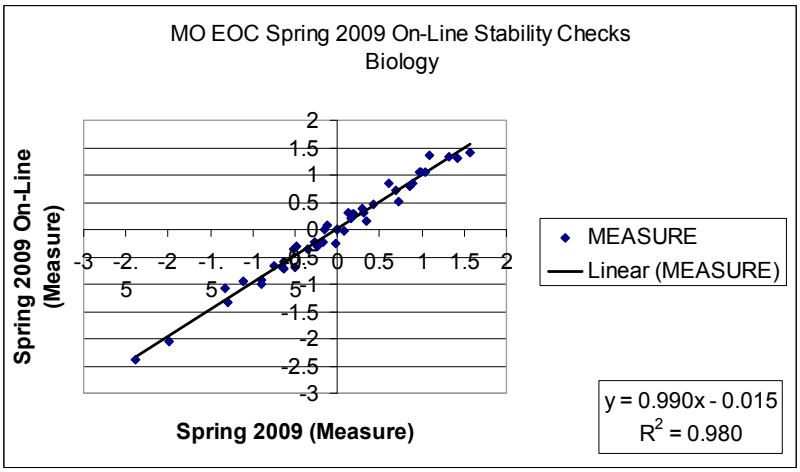


Figure 3. Scatterplot of the Common Items Between Online and P/P Administrations: Biology



A final fixed calibration was performed on the online data after removing the unstable item(s) from the common set of anchor items. Finally, raw score-to-scale score conversion tables were created using Rasch scaling with Winsteps software.

Tables 23 through 25 present the side-by-side raw score-to-scale score conversions for the paper/pencil and online "forms." Note that the raw score-to-scale score conversions for the online administration have not been used operationally. However, these analyses are included to represent the impact on the score scale should a decision be made to "equate" the online administrations onto the paper/pencil scale.

Copyright © by Missouri Department of Elementary and Secondary Education

Table 23: Differences between Scale Scores for Online Matched Sample–English II

Raw Score	Associated Scale Score for Paper/Pencil	Associated Scale Score for Online	Paper/Pencil Minus Online Scale Score
0	100	100	0
1	105	105	0
2	125	126	-1
3	137	138	-1
4	145	145	0
5	150	151	-1
6	155	155	0
7	158	159	-1
8	162	162	0
9	165	165	0
10	168	168	0
11	171	171	0
12	173	173	0
13	176	176	0
14	178	178	0
15	180*	180*	0
16	182	183	-1
17	185	185	0
18	187	187	0
19	189	189	0
20	191	191	0
21	193	193	0
22	196	196	0
23	198	198	0
24	200**	200**	0
25	202	202	0
26	205	205	0
27	207	207	0
28	210	210	0
29	212	212	0
30	215	215	0
31	218	218	0
32	221	221	0
33	225***	225***	0
34	229	229	0
35	234	234	0
36	240	240	0
37	248	248	0
38	250	250	0
39	250	250	0

* Basic cut.
** Proficient cut.
*** Advanced cut.

Copyright © by Missouri Department of Elementary and Secondary Education

Table 24: Difference between Scale Scores for Online Matched Sample–Algebra I

Raw Score	Associated Scale Score for Paper/Pencil	Associated Scale Score for Online	Paper/Pencil Minus Online Scale Score
0	100	100	0
1	112	112	0
2	127	127	0
3	136	136	0
4	143	143	0
5	149	149	0
6	154	154	0
7	158	158	0
8	162	162	0
9	165	166	-1
10	169	169	0
11	172	172	0
12	177*	177 (175)*	0
13	178	178	0
14	181	181	0
15	183	183	0
16	186	186	0
17	189	189	0
18	191	191	0
19	194	194	0
20	196	197	-1
21	200**	200 (199)**	0
22	202	202	0
23	204	204	0
24	207	207	0
25	210	210	0
26	212	213	-1
27	215	215	0
28	218	218	0
29	221	221	0
30*	225***	225***	0
31	228	228	0
32	232	232	0
33	236	237	-1
34	242	242	0
35	249	248	1
36	250	250	0
37	250	250	0
38	250	250	0

Notes: The paper/pencil calibration was run on 38 items, so the raw score-to-scale score conversion is not the same as the operational raw score-to-scale score tables used for reporting. Raw score cuts were reduced by one score point from the operational test because item #26 was dropped from the analysis.
* Basic cut.
** Proficient cut.
*** Advanced cut.

Copyright © by Missouri Department of Elementary and Secondary Education

Table 25: Difference between Scale Scores for Online Matched Sample–Biology

Raw Score	Associated Scale Score for Paper/Pencil	Associated Scale Score for Online	Paper/Pencil Minus Online Scale Score
0	100	100	0
1	107	107	0
2	121	121	0
3	130	130	0
4	137	136	1
5	142	142	0
6	146	146	0
7	150	150	0
8	153	153	0
9	156	156	0
10	159	159	0
11	162	162	0
12	164	164	0
13	166	166	0
14	169	168	1
15	171	171	0
16	173	173	0
17	175	174	1
18	177	176	1
19	178	178	0
20	180*	180*	0
21	182	182	0
22	184	183	1
23	185	185	0
24	187	187	0
25	189	188	1
26	190	190	0
27	192	192	0
28	193	193	0
29	195	195	0
30	197	197	0
31	198	198	0
32	200**	200**	0
33	202	202	0
34	203	203	0
35	205	205	0
36	207	207	0
37	208	208	0
38	210	210	0
39	212	212	0
40	214	214	0
41	216	216	0
42	218	218	0
43	220	220	0
44	223	222	1

Copyright © by Missouri Department of Elementary and Secondary Education

Table 25: Difference between Scale Scores for Online Matched Sample–Biology (continued)

45	225***	225***	0
46	228	228	0
47	231	230	1
48	234	234	0
49	238	237	1
50	242	242	0
51	247	247	0
52	250	250	0
53	250	250	0
54	250	250	0
55	250	250	0

* Basic cut.
** Proficient cut.
*** Advanced cut.

Differential Item Functioning Analysis

Differential item functioning (DIF) was used to further study the possible effects of mode on student item responses. DIF was examined with the Mantel-Haenszel (1959) procedure for the SR items and with a Rasch DIF analysis using Winsteps (v3.64, Linacre, 2006) for the PE items. The Mantel-Haenszel (MH) method is a nonparametric approach to DIF utilizing chi-square and delta statistics to classify each item. The classification system, as well as the computational formulas for Mantel-Haenszel and delta statistics, are described in detail elsewhere (e.g., Camilli and Shepard 1994; Dorans and Holland 1993). Items classified as "A" are considered to exhibit a negligible amount of DIF, "B" items to exhibit a moderate amount of DIF, while "C" items exhibit a large amount of DIF.

DIF detection with Winsteps is a Rasch-model-based approach. The Rasch and Mantel-Haenszel procedures for DIF are equivalent under certain conditions (Linacre and Wright, 1989; Schulz, Perlman, Rice, and Wright, 1996). Similar to the ETS classifications, the DIF output yielded by Winsteps is classified as negligible (A), slight to moderate (B), or moderate to severe (C). If a *t*-value is smaller than 2.58 or the DIF contrast is smaller than 0.45 logits, the item is flagged as A. If a *t*-value is larger than 2.58 and the DIF contrast is larger than 0.65 logits, the item is flagged as C. Otherwise, items are flagged as B.

In the DIF analysis, students who took the P/P test served as the reference group, while students who took the online test served as the focal group. Items exhibiting a moderate to large amount of DIF might be functioning differently between the two modes of administration (i.e., the items may be easier for students taking the test via one or the other mode of administration). Summaries of the results of the analyses for both the total sample and matched sample are presented in Table 26. Note that there was little evidence of DIF when either the total samples or the matched samples were examined.

Copyright © by Missouri Department of Elementary and Secondary Education

Table 26: Frequency Distribution of DIF Categories by Item Type for Total Sample and Matched Sample

Sample/ Assessment	Selected Response (SR) Items*						Performance Events (PE) Items*					
	A**	A–**	B**	B–**	C**	C–**	A**	A–**	B**	B–**	C**	C–**
Total												
English II	34	0	0	1	0	0	1	0	0	0	0	0
Algebra I	34	0	0	1	0	0	1	0	0	0	0	0
Biology	35	0	0	0	0	0	11	0	0	0	0	0
Matched												
English II	34	0	0	1	0	0	1	0	0	0	0	0
Algebra I	35	0	0	0	0	0	1	0	0	0	0	0
Biology	35	0	0	0	0	0	10	0	1	0	0	0

Note: Classifications with a negative sign ("–") favor the reference group, while classifications with no sign favor the focal group. DIF categories: A, negligible; B, slight to moderate; and C, moderate to severe.
* The Mantel-Haenszel procedure is applied for the SR items and Rasch DIF analysis for the PE/WP items.

Factor Analysis for Matched Sample

Both exploratory and confirmatory factor analyses were conducted using the matched sample. Exploratory analyses were conducted first to establish general factor models and evaluate the comparability of the paper/pencil and on-line solutions. The factor patterns for the P/P group were then input as fixed for confirmatory factor analyses to more precisely evaluate the congruence of the factor structures for the two groups.

Exploratory Factor Analyses

English II. The extraction method utilized was Iterated Principal Factors with an Oblique Varimax rotation. The initial runs with operational data yielded 5 eigenvalues that were greater than 1. Upon review of the initial eigenvalue scree plot as well as the difference between reduced correlation matrix eigenvalues (successive eigenvalues showed little change after the fourth factor), and factor loading interpretability issues, it was decided to retain 4 factors.

Algebra I. The extraction method utilized was Iterated Principal Factors with an Oblique Varimax rotation. The initial runs with operational data yielded 5 eigenvalues that were greater than 1. Upon review of the initial eigenvalue scree plot as well as the difference between reduced correlation matrix eigenvalues (successive eigenvalues showed little change after the fourth factor), and factor loading interpretability issues, it was decided to retain 4 factors.

Biology. The extraction method utilized was Iterated Principal Factors with an Oblique Equamax rotation. The initial run with the operational data yielded 13 eigenvalues that were greater than 1. After review of the initial eigenvalue scree plot, differences between reduced correlation matrix eigenvalues (successive eigenvalues showed little change after the fourth factor), and factor loading interpretability issues, it was decided to retain 4 factors.

Copyright © by Missouri Department of Elementary and Secondary Education

Table 27. Factor Analysis of Tetrachoric Correlation Coefficients for the Operational Items: English II

Item	Factor 1	Factor 2	Factor 3	Factor 4
Item 1	58*	41*	47*	35
Item 2	42*	33	22	22
Item 3	46*	25	16	15
Item 4	79*	51*	29	36
Item 5	54*	39	29	33
Item 6	8	10	36	17
Item 7	50*	40*	45*	37
Item 8	59*	44*	37	35
Item 9	28	25	50*	21
Item 10	47*	38	31	30
Item 11	72*	58*	40*	43*
Item 12	36	31	37	25
Item 24	32	30	36	25
Item 25	32	32	42*	26
Item 26	59*	55*	44*	44*
Item 27	60*	53*	61*	36
Item 28	32	31	27	21
Item 29	16	20	2	13
Item 30	34	38	54*	31
Item 31	43*	53*	38	40*
Item 32	53*	63*	52*	45*
Item 33	24	26	44*	26
Item 34	34	41*	35	30
Item 35	46*	62*	32	49*
Item 36	48*	84*	31	53*
Item 37	28	45*	28	24
Item 38	34	41*	39	29
Item 39	38	65*	26	52*
Item 40	30	45*	33	36
Item 41	34	53*	24	44*
Item 43	35	49*	31	57*
Item 44	34	33	41*	41*
Item 45	19	31	14	63*
Item 46	11	25	18	54*
Item 47	34	33	33	51*

Notes: Factor loading values are multiplied by 100 and rounded to the nearest integer. Factor loadings greater than 0.396931 are flagged by an asterisk. The flag criterion value of 0.396931 is the root mean square of all of the values in the matrix.

Copyright © by Missouri Department of Elementary and Secondary Education

Table 28. Factor Analysis of Tetrachoric Correlation Coefficients for the Operational Items: Algebra I

Item	Factor 1	Factor 2	Factor 3	Factor 4
Item 1	45*	44*	22	30
Item 2	21	26	12	20
Item 3	47*	47*	51*	24
Item 4	35	33	25	23
Item 5	40*	57*	58*	31
Item 10	24	44*	39	37
Item 11	42*	40*	27	22
Item 12	49*	47*	44*	25
Item 13	49*	38	39*	46*
Item 14	57*	21	33	38
Item 15	67*	47*	40*	34
Item 16	56*	47*	39*	35
Item 17	47*	49*	35	39*
Item 18	41*	63*	44*	33
Item 19	61*	60*	51*	43*
Item 20	47*	27	26	34
Item 21	58*	34	42*	41*
Item 27	33	44*	43*	30
Item 28	44*	35	44*	41*
Item 29	27	23	57*	38
Item 30	25	38	42*	35
Item 31	31	25	29	33
Item 32	36	32	50*	29
Item 33	56*	28	45*	38
Item 34	27	15	62*	38
dropped				
Item 36	24	27	41*	54*
Item 37	39*	15	45*	50*
Item 38	40*	41*	46*	22
Item 43	18	24	34	26
Item 44	38	28	35	30
Item 45	42*	26	22	37
Item 46	15	18	15	34
Item 47	22	-1	20	39*
Item 48	26	21	42*	30

Notes: Factor loading values are multiplied by 100 and rounded to the nearest integer. Factor loadings greater than 0.387648 are flagged by an asterisk. The flag criterion value of 0.387648 is the root mean square of all of the values in the matrix.

Copyright © by Missouri Department of Elementary and Secondary Education

Table 29. Factor Analysis of Tetrachoric Correlation Coefficients for the Operational Items: Biology

Item	Factor 1	Factor 2	Factor 3	Factor 4
Item 1	39 *	38	30	54 *
Item 2	51 *	47 *	45 *	51 *
Item 3	16	20	10	27
Item 4	41 *	37	37	15
Item 5	26	38	23	37
Item 10	36	40 *	34	48 *
Item 11	27	32	36	46 *
Item 12	34	24	39 *	20
Item 13	46 *	32	54 *	41 *
Item 14	30	32	34	28
Item 15	19	24	26	24
Item 16	30	34	37	53 *
Item 17	41 *	39	25	64 *
Item 18	39	38	48 *	59 *
Item 19	41 *	43 *	50 *	45 *
Item 20	25	12	33	10
Item 21	20	20	26	27
Item 26	43 *	43 *	50 *	40 *
Item 27	38	34	38	37
Item 28	60 *	58 *	59 *	48 *
Item 29	33	30	52 *	30
Item 30	19	20	27	31
Item 31	36	28	51 *	42 *
Item 32	41 *	41 *	44 *	52 *
Item 33	45 *	39 *	54 *	41 *
Item 34	44 *	38	56 *	33
Item 35	23	26	22	49 *
Item 36	35	43 *	42 *	31
Item 37	36	29	46 *	25
Item 38	34	30	32	41 *
Item 43	19	20	19	30
Item 44	38	40 *	47 *	24
Item 45	37	36	53 *	38
Item 46	31	34	45 *	31
Item 47	51 *	43 *	48 *	44 *
Item 48	33	53 *	18	34
Item 49	91 *	31	18	14
Item 50	92 *	37	20	14
Item 51	49 *	60 *	48 *	11
Item 52	40 *	38	28	28
Item 53	27	35	24	26
Item 54	41 *	55 *	33	38
Item 55	33	62 *	18	18
Item 56	42 *	77 *	20	42 *
Item 57	33	49 *	29	34
Item 58	22	37	14	11

Notes: Factor loading values are multiplied by 100 and rounded to the nearest integer. Factor loadings greater than 0.386801 are flagged by an asterisk. The flag criterion value of 0.386801 is the root mean square of all of the values in the matrix.

Copyright © by Missouri Department of Elementary and Secondary Education

Confirmatory Factor Analysis

The purpose of the confirmatory analyses was to evaluate the similarity of the factor structures across the online and P/P groups. Our approach represents somewhat of a departure from typical applications of confirmatory factor analysis. In the current analyses, we were primarily concerned not with model fit per se, but with the comparability of model fit between the P/P and online data. In the analyses reported below, the factor and factor correlation patterns as well as the specific values of the loadings and correlations were constrained. Thus, loadings and factor correlations were not estimated; rather, our interest was in the remaining residual item covariances and model fit under a fully constrained model, and more specifically in the similarity of the fit and residual covariances between the online and paper/pencil groups.

The analyses were carried out using a nonlinear factor analysis approach, as implemented by the program NOHARM (Fraser, 1988). The program computes the residual covariances of the items, after fitting the model, and gives the root mean square of the residual covariances as an overall measure of misfit of the model to the data, with the given number of dimensions and the given pattern of the coefficient matrix.

Input to the program consisted of the factor loadings and factor correlations obtained from the P/P exploratory factor analyses. The confirmatory analyses were run for both the online and P/P data sets, with the latter providing the baseline for comparison of the fit statistics. Fit and similarity of fit was evaluated using the sum of squared residual covariances (SSR) and the root mean squared residuals (RMSR) for each dataset, the mean absolute difference of residuals between data sets (MARD), and the matrix of absolute residual differences (ARD). An examination of the residual difference matrix for clusters of large residual differences may indicate items that do not fit the model similarly between the two groups.

Results of Confirmatory Factor Analyses

Global Model Fit. Table 30 provides the results of the summary fit indices. For English II and Algebra I, the SSR and RMSR values were somewhat higher for the online data. For Biology, the SSR and RMSR were slightly higher for the P/P data. According to Fraser (1988), if the root mean square residual is in the order of the typical standard error of the residuals (4 times the reciprocal of the square root of the sample size) we have a rough indication that a refined test of significance would not reject the hypothesized model. With the sample sizes involved in these analyses. the RMSR values in Table 30 are all well below their approximate standard errors, and therefore the constrained models obtained from the exploratory analyses appear to fit the data well, and do so for both the P/P and on-line data.

Also included in Table 30 is the Tanaka (1993) unweighted least squares goodness-of-fit index. There are no interpretative guidelines for Tanaka's index, other than in general a higher value implies better model fit. The values in Table 30 are high for all analyses. Consistent with the SSR and RMSR values, the Tanaka index indicates slightly better fit for the P/P data for Algebra I and English I, and for the on-line data for Biology.

Finally, the mean absolute residual difference (MARD) reflects the average absolute discrepancy in residual covariances between the online and paper/pencil data. The values

Copyright © by Missouri Department of Elementary and Secondary Education

in Table 30 are low; indicating that, on average, the differences in misfit between the two data sets was small. This is explored in more detail below.

Table 30: Summary of Fit Statistics from the Confirmatory Factor Analysis

	Mode of Administration	SSR	RMSR	Tanaka GFI	MARD (Online-P/P)
English II	P&P	.0257	.0066	.976	
	On-line	.0465	.0088	.957	.0043
Algebra I	P&P	.0497	.0091	.966	
	On-line	.0629	.0103	.955	.0048
Biology	P&P	.1289	.0112	.939	
	On-line	.1186	.0107	.944	.0070

Evaluation of Item Residual Covariances. Table 31 contains the frequencies of absolute residual differences (ARD) across the items in each test. As can be seen, the majority of the ARDs are below .01. The greatest similarity in residuals occurred for English II (93.95% below .01) followed by Algebra II (87.4%) and Biology (75.1%).

Table 31: Frequency Distribution of Absolute Residual Differences (ARD)*

	English II *N* (%)	Algebra I *N* (%)	Biology *N* (%)
ARD < .01	559 (93.95%)	520 (87.4%)	797 (75.1%)
ARD ≥ .01	36 (6.05%)	75 (12.6%)	238 (24.9%)

* Lower diagonal matrices

Tables 32 through 34 describe the contents of the residual difference (ARD) matrices for English II, Algebra I and Biology, respectively. Specifically, the tables report the number of items by frequency of occurrence of ARD ≥ .01 and the percent of total. This is computed using the full residual difference matrices. These tables provide a summary of the dispersion (or concentration) of ARD ≥ .01 across items. For example, for English II, there are 2 items associated with 8 occurrences of ARD ≥ .01 and 1 item with six occurrences of ARD ≥ .01. In the full residual differences matrix there are 72 total occurrences of ARD ≥ .01 so that these three items were associated with 30% of these occurrences. For Algebra I, it can be seen that two items were associated with 31 (12 + 19) of the 150 total occurrences, or 21%.

For Biology, there were several items associated with multiple occurrences of ARD ≥ .01, and the misfit and misfit differences were fairly spread out throughout the matrix. Three items had occurrences of at least 20, but given the spread and larger size of this matrix, these items did not account for a large proportion of the residual differences as was the case for the other tests.

Copyright © by Missouri Department of Elementary and Secondary Education

Table 32: Number of Items by Frequency of Occurrence of ARD ≥ .01: English II

Number of Items with ARD ≥ .01	Frequencies of ARD ≥ .01	Percent of Total ARD≥ .01*
9	1	.13
5	2	.14
5	3	.21
4	4	.22
1	6	.08
2	8	.22
Total: 26		

*** Full residual difference matrix**

Table 33: Number of Items by Frequency of Occurrence of ARD ≥ .01: Algebra I

Number of Items with ARD ≥ .01	Frequencies of ARD ≥ .01	Percent of Total ARD≥ .01*
1	1	.01
5	2	.07
11	3	.22
4	4	.11
4	5	.13
4	6	.16
1	7	.05
1	8	.05
1	12	.08
1	19	.13
Total: 34		

*** Full residual difference matrix**

Copyright © by Missouri Department of Elementary and Secondary Education

Table 34: Number of Items by Frequency of Occurrence of ARD ≥ .01: Biology

Number of Items with ARD ≥ .01	Frequencies of ARD ≥ .01	Percent of Total ARD≥ .01*
3	1	.01
2	6	.03
1	7	.01
3	8	.05
4	9	.08
7	10	.15
2	11	.05
5	12	.13
5	13	.14
7	14	.21
2	15	.06
1	17	.04
2	20	.08
1	22	.05
Total: 45		

*** Full residual difference matrix**

Discussion

The confirmatory analyses were run to evaluate the relative fit of P/P versus online data to the same imposed factor model. The results were mixed. For English II and Algebra I, the fit to the four-factor model was good, and slight better globally for the P/P data. At the item level, most residual covariances as well as differences in residuals were very small. For both tests there were one or two items that accounted for a disproportionate amount of the differences in item misfit.

For Biology, the model fit was weaker in general but still good, and in contrast to that for Algebra I and English II, slightly better for the on-line data. Differences in item misfit were rather widely dispersed.

These results indicate that there may be a mode effect for Biology, but there are little or no indications of a general effect between the online and P/P data for English II and Algebra I. There may be two or three items in each of those tests that behave somewhat differently, in terms of not fitting the factor model identically, but for most of the items in both datasets the fit was very good. Also, it is important to emphasize that the global fit indices indicated good and nearly comparable fit for all three tests in both modes.

Copyright © by Missouri Department of Elementary and Secondary Education

Summary and Conclusions

If the results of the confirmatory factor analysis tend to support an assumption that the factor structures of the two administration modes are comparable, what does that mean for an overall decision of comparability? This result seems suggestive, but given the apparent differences in the P/P and online samples, it does not seem sufficient for a determination that there is no mode effect. Clearly the P/P and the voluntary online samples were not comparable, and this result confounds our ability to interpret differences in student performance across the modes. Efforts were made, both by building matched samples and by applying analysis of covariance, to remove some of the demographic differences between the samples post hoc, and the performance of the two groups did seem to move closer together. However, there still were small differences for English II and Algebra. For these samples, the P/P group tended to perform slightly better than the online group. Sufficient information may still not exist to completely disentangle the effects of mode and the non-random effects of sample selection. However, the evidence that has been obtained suggest that 1) there is little appreciable difference in the factor structures of the tests delivered by the different modes; 2) there is little evidence of DIF to indicate that certain types of items function differently by mode; and 3) there *do* seem to be differences in mean performance by mode for the Algebra and English II tests, but the differences seem to be practically small.

Copyright © by Missouri Department of Elementary and Secondary Education

References

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

Fraser, C. (1988). NOHARM: *An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, Australia: The University of New England. Froelich, A. G. (2000).

Linacre. J.M., 2006. Winsteps 3.64 [Computer software program]. Chicago: Author.

Linacre, J. M., and B. D. Wright. 1989. The equivalence of Rasch, PROX, and Mantel-Haenszel. *Transactions of the Rasch Measurement Special Interest Group* 3:1–3.

Mantel, N., and W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute* 22:719–748.

Schulz, E. M., C. Perlman, W. K. Rice, and B. D. Wright. 1996. An empirical comparison of Rasch and Mantel-Haenszel procedures for assessing differential item functioning. In *Objective measurement: Theory into practice*, Vol. 3, ed. G. Engelhard and M. Wilson, 65–84. Norwood, NJ: Ablex Publishing.

Tanaka, J. S. (1993). Multifaceted concepts of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 10–39). Newbury Park, CA: Sage.

Copyright © by Missouri Department of Elementary and Secondary Education