Comparability Study of Online and Paper and Pencil Tests Using Modified Internally and Externally Matched Criteria

Thakur Karkee
Measurement Incorporated

Dong-In Kim
CTB/McGraw-Hill

Kevin Fatica
CTB/McGraw-Hill

Paper presented at the annual meeting of the
American Educational Research Association (AERA)
Denver, CO
April 29 – May 4, 2010

Abstract

Many states are exploring possibilities of transitioning to online mode of test administration with the premise that future tests will be computer based of some sort. The benefits include quick on-demand reporting, immediate feedback, and cost effectiveness. Comparability studies are conducted to help demonstrate and ensure the defensibility of using the scores interchangeably between paper and pencil and online tests. Common designs used in comparability study require either random assignment of examinees or assign testing mode to random equivalent groups.

Alternate designs conditioned on internal and internal plus external criteria for creating equivalent samples are compared in this study together with the scores between modes of administration with in a condition. The mode effects at item and student level were evaluated by comparing model fit, differential item functioning, and mean of item and person parameters. The test results did not show statistically discernible mode effects based on model fit, DIF, or student performance, despite some differences in item parameters.

Key words: Comparability Study, alternate designs; online and paper and pencil tests; internal and external matching.

Introduction

Many state testing programs are exploring the possibility of transitioning from a paper and pencil mode of test administration to an online mode. Changing the mode of administration creates questions as to whether or not the tests, and test scores derived from those tests, are equivalent, and whether there are effects of the testing mode. Comparability studies are conducted to determine if such mode effects exist, and if scores derived from the same test in two different modes are comparable and can be used interchangeably.

Paek (2005) reviewed the current research trends in comparability studies and indicated that the research on mode effects shows mixed results. Some studies have shown presence of a mode effect for students tested online such that their performance was lower because of scrolling requirements for passage-associated items (Poggio, Glassnapp, Yang, & Poggio, 2005; Way, Davis & Fitzpatrick, 2006) or items that require graphing (Ito & Sykes, 2004; Keng, McClarty & Davis, 2006). On the other hand, Mead and Drasgow (1993) in their meta-analyses of comparability studies concluded that there were essentially no mode effects for the power tests they analyzed. Similarly, Wang (2004) found no mode effects for the Stanford Diagnostic Reading and Mathematics tests, although there were some exceptions.

Most comparability studies used variety of randomization methods to create equivalent samples of students in the online and paper and pencil testing modes, which then enabled the researchers to assess mode effect across modes of test administration. The two most common comparability study designs are the *randomized equivalent group design* (REGD), and the *single-group repeated measure design* (SGRMD). In the REGD approach, examinees are randomly assigned to either paper and pencil or to online testing, and then, on the basis of the notion that the randomly created groups are equivalent, differences in the scores of the testing groups are attributed to a mode effect. In the SGRMD, a single group of students takes the test in both modes. This design allows for the comparison of student ability in both modes.

These two common designs have advantages and shortcomings. While REGD has an advantage over SGRMD in that students take only one test, REGD requires a large number of students to be randomly assigned to each of the two modes, and this random assignment of mode to individual students within schools is a challenging task that is difficult to control. SGRMD has advantages in that it reduces experimental error and increases statistical power in detecting a mode effect by controlling for individual variation in performance, but SGRMD requires that students test twice, which can create practical test administration and motivational problems.

In part because these designs have shortcomings, researchers continue to investigate additional approaches to comparability studies. One of the more interesting approaches taken recently is addressed in two studies by Way et al (2006, 2007). The current paper extends the approach developed there and advances some questions raised in that research.

The Way et. al. (2006, 2007) studies present two alternative ways of creating equivalent groups of examinees across testing modes that do not require randomization. Conceptually, like the REGD approach, the logic of the designs discussed by Way et. al. uses the equivalence of the groups compared across testing modes as the basis for attributing differences in testing results across groups to the testing mode. In the first case, Way et. al. (2006) created equivalent samples of students in a Reading and Math test in grade 8 based on their performance on the Reading and Math tests in grade 7. Predicted scores for the grade 8 test, independent of testing mode, were established on the basis of the grade 7 test, using regression. The samples of examinees were "equivalent" in terms of their predicted scores and demographic variables. Way et. al. (2007) modified this approach by adding different content area in the prediction equation, such as, for example, using Reading scores in addition to Science scores in a grade 7 test to predict scores on a Science test in grade 8. The design which builds equivalent samples of students based on a predicted score within a single content area is referred to as using an *internal* criterion. Using scores from one content area to predict scores in another content area is referred to as using an *external* criterion.

The current study extends the Way et. al. (2007) study by conditioning samples on the internal criteria and the internal plus external criteria simultaneously. In addition, the current study also offers some insight into designing comparability studies for different circumstances: the current study uses testing data from a recent state-level testing program over two administrations of an end-of-instruction (EOI) Social Studies test. In the first administration, all testing was in a paper and pencil mode. In the second year, about half of the schools in the state moved to online testing, but the other half continued with the paper and pencil mode. The testing mode, in other words, was assigned at the school level not at the individual student level. In addition, the schools were not assigned to one of the modes randomly, the participation was voluntary.

As noted, the data used in this study come from an EOI test. This situation presents a different set of circumstances than those examined in previous comparability studies in that there were no comparable prior scores for the same students from prior grades to use as a basis for predicting performance in the EOI test, nor was it possible to select and create groups of students to compare in a subsequent year after the EOI test. The design developed to deal with these circumstances, as described further below, uses the performance of one cohort (in year 1) to predict the performance of the next cohort (in year 2), when the online mode of administration was introduced. This design adaptation to an EOI test is a significant departure from the approach used by Way et. al. (2006, 2007) and is unique in the literature. The current study thus extends the line of inquiry taken up in the Way, et. al. studies by expanding the basic approach to different and unique circumstances and by asking two basic research questions: First, does using both *internal* and *external* criteria to establish equivalent samples offer any advantage over using only *internal* criteria? And second, provided that the two samples are equivalent based on the school-level information, does a mode effect exist in the EOI test?

Methods

Study Design

The study design is shown in Table 1. Four samples were selected for this study. Sample 1 (S1) and Sample 2 (S2), are matched to the prior paper and pencil (PP) administration based on internal criteria. Sample 3 (S3) and Sample 4 (S4) are matched to the prior PP administration based on both internal matched (IM) and internal plus external matched (IEM) criteria. In the IM condition, S1 is pulled from the online (OL) schools in year 2, and S2 is pulled from the PP schools in year 2. In the IEM condition, S3 is pulled from the OL schools in year 2, and S4 is pulled from the PP schools in year 2. By design, some schools in the OL (S1 and S3) and PP (S2 and S4) samples are common to the IM and IEM conditions.

Table 1. Comparability Study Design

Matched Criteria	Year 1 Mode	Year 2 Mode	
		Online	Paper and Pencil
Internal	Paper-and-Pencil	Sample 1 (S1)	Sample 2 (S2)
Internal +External	Paper-and-Pencil	Sample 3 (S3)	Sample 4 (S4)

Data

The data for this study came from a recent large scale end-of-instruction statewide test in Social Studies. All test items were multiple-choice (MC). A total of over 50,000 students participated in each of the two administrations from which data are drawn. In year 2 of the study, approximately 50% of the schools in the state voluntarily chose to administer an online version of the test, while the other half of the school continued with the paper and pencil version. As a result, there were approximately 25,000 students in each testing mode in year 2. The testing program used item response theory (IRT) to calibrate, scale, and score student responses.

Creating Equivalent Samples

As noted, the approach used here is similar to that discussed by Way et. al. (2006, 2007). The logic of this design revolves around using the equivalence of groups compared across testing modes as the basis for attributing differences in testing results across groups to the testing modes. Both the internal and the external matching are based on regression. Both approaches use results from year 1 to establish predictions for performance in year 2.

The internally matched samples use school-level performance scores and demographic data from the Social Studies EOI test in year 1, where all testing was in PP mode, to predict scores in the Social Studies EOI test in year 2, where testing occurred in both OL and PP modes. The internally plus externally matched (IEM) samples use results from the Social Studies EOI test in year 1, as well as the Reading Language Arts (RLA) test in year 1, to predict scores in the Social Studies EOI test, in year 2.

As a first step in creating the IM sample, the year 1 ability estimates based on the PP test were obtained from the administration's database for all students in all schools. The next step was to establish school-level ability estimates for each school in year 2, based on all student responses from both testing modes. To do this, all student responses from both testing modes were combined together, calibrated using three-parameter logistic (3PL) IRT model, and then scaled. The estimated item parameters were used to score all student responses and to estimate each student's ability (theta). School-level thetas were then calculated as the average of the student thetas in the school. The schools in year 1 and year 2 were then matched using a unique school identification code. The observed school-level thetas from the two administrations were then regressed to obtain a predicted school-level theta for each school in year 2. A linear regression model was applied, using the year 2 school-level mean theta as the dependent variable and the year 1 school-level mean theta as the independent variable. The regression equation is shown below:

$$\hat{X}_{i2} = \beta_0 + \beta_1 X_{i1}$$

where $\hat{X}_{i2}$ is the predicted school-level theta for content X on the EOI test for ith school in year 2, X_{i1} is the observed score for the ith school in the content in year 1, and β_0 and β_1 are regression coefficients. The predicted school-level thetas for year 2 from each of the schools in the total sample, irrespective of mode, were then sorted into 10 theta intervals, ranging from low to high ability. Then, we identified the OL and PP school samples within each of the 10 theta intervals in such a way that the weighted mean theta and demographic distributions (based on gender and ethnicity) matched as closely as possible. This procedure provided a total sample size of approximately 6000 cases in each sample. Note that there were more schools in the middle of the theta distribution and fewer at the extremes. Therefore, the OL and PP samples from the middle theta intervals included proportionally more schools, and ultimately more cases, than from the extreme intervals. After identifying OL and PP schools at each theta interval as described, we then pulled the student response data from the selected schools to create equivalent mode-specific samples.

The procedures used to create the IM samples were replicated to create the IEM samples as well. The only difference was that the school-level mean theta for the RLA content area was added as an additional criterion for selecting the equivalent samples. In this case, the predicted mean thetas for year 2 were based on both the Social Studies and RLA mean school-level thetas from the year 1 administration. Mathematically, we applied a multiple regression to obtain the predicted mean school theta:

$$\hat{X}_{i2} = \beta_0 + \beta_1 X_{i1} + \beta_2 Z_{i1}$$

where $\hat{X}_{i2}$ is the year 2 predicted ith school-level theta score in the EOI test; X_{i1} is the year 1 observed score in the same content area; Z_{i1} is the year 1 observed score for the RLA content; and $\beta_0,$ $\beta_1,$ $\beta_2,$ are multiple regression coefficients. As in the internally matched method, the predicted school-level thetas were then divided into 10 equal intervals, independent of testing mode. Two equivalent samples were then drawn from OL schools and PP schools at each theta interval by matching mean theta and demographic profiles in such a way that the total sample size comprised approximately 6000 cases in each sample.

The sample characteristics based on the internal matching are shown in Table 2 and on the internal plus external matching are shown in Table 3. These tables show the theta interval, testing mode, the number of students, the mean theta, and gender and ethnicity distributions for each sample group. The mean theta and percentages of students in the gender and ethnicity distributions for a given theta interval and testing mode are the weighted average of the corresponding number of students in the sample schools in the interval.

Table 2 reflects the approximate normal distribution of the mean theta with respect to the number of students (N). There are more students in the middle of the theta range (intervals 4, 5, and 6) and fewer students towards the extremes (intervals 3, 7, 8, and 9). These numbers are the approximate proportional representation of the total number of schools and students in these intervals. No sample schools are selected from theta intervals 1, 2, and 10 as no schools that took the test in the online mode were in theta intervals 1 and 10, and no schools that took the test in the paper and pencil mode were in interval 2.

Table 2. Mean Theta and Demographic Profiles for Internally Matched Samples

Theta Interval	Mode	N	Mean Theta	Male	Female	American Indian	Asian	African American	Hispanic	White (non-Hispanic)
3	OL	263	-0.58	47.53	52.47	25.10	1.14	2.66	4.18	65.02
3	PP	236	-0.58	47.03	52.97	30.51	2.12	0.42	4.66	61.86
4	OL	1870	-0.34	48.82	51.18	15.72	3.42	13.05	5.51	61.55
4	PP	1898	-0.33	52.05	47.95	15.07	4.16	15.33	3.00	60.96
5	OL	1424	-0.11	52.46	47.54	20.93	2.04	14.40	5.27	56.46
5	PP	1109	-0.11	50.32	49.68	32.46	0.81	4.51	3.79	57.17
6	OL	1843	0.12	48.07	49.01	22.41	1.82	10.39	5.35	56.43
6	PP	1855	0.11	51.70	48.30	17.25	3.02	13.42	5.77	58.65
7	OL	263	0.38	49.43	50.57	22.81	0.00	2.28	5.70	69.20
7	PP	215	0.35	47.91	52.09	15.81	3.25	6.51	6.05	65.11
8	OL	137	0.64	46.71	53.29	35.03	0.00	0.73	5.84	58.40
8	PP	494	0.64	50.61	49.39	15.79	1.01	25.51	7.29	48.58
9	OL	74	0.87	43.24	56.76	28.38	0.00	12.16	9.46	50.00
9	PP	78	0.82	52.57	47.43	14.10	0.00	0.00	2.57	82.05

Table 2 also shows that, within each theta level, the gender and ethnicity distributions for each testing mode are similar. For example, “Theta Interval 3” consisted of 263 OL students and 236 PP students with a weighted mean theta of -0.58 in both the OL and PP testing modes. The male/female ratio was 47.53 to 52.47 in OL mode and 47.03 to 52.97 in PP mode. The ethnicity distribution was also similar in the OL and the PP samples. A similar pattern can be observed in the sample groups in both modes for all theta intervals. The same basic interpretation can be applied to Table 3, which shows the samples created based on internal plus external matching.

Table 3. Mean Theta and Demographic Profiles for Internally plus Externally Matched Samples

Theta Interval	Mode	N	Mean Theta	Male	Female	American Indian	Asian	African American	Hispanic	White (non-Hispanic)
4	OL	174	-0.91	50.00	50.00	24.71	4.02	4.60	6.32	60.34
4	PP	274	-0.91	51.09	48.91	25.18	4.02	6.57	5.11	58.76
5	OL	572	-0.70	50.35	49.65	31.29	0.35	19.93	1.75	46.33
5	PP	594	-0.70	50.34	49.66	14.31	5.05	20.54	15.99	43.26
6	OL	1542	-0.45	50.65	49.35	25.29	1.49	13.04	4.22	55.97
6	PP	1495	-0.47	46.36	53.64	27.83	1.34	7.56	3.95	57.73
7	OL	1253	-0.25	49.64	50.36	22.98	0.72	6.22	4.63	63.77
7	PP	1298	-0.25	50.39	49.61	23.19	2.16	8.48	3.24	60.63
8	OL	1208	-0.04	51.16	48.84	16.47	2.81	8.28	6.54	65.31
8	PP	1184	-0.02	50.42	49.58	10.56	4.73	12.33	4.90	64.53
9	OL	1034	0.20	48.07	51.93	12.57	4.84	8.22	5.70	67.70
9	PP	932	0.20	53.32	46.68	10.52	3.11	20.17	3.01	62.77
10	OL	100	0.36	46.00	54.00	16.00	3.00	7.00	6.00	64.00
10	PP	124	0.46	52.42	47.58	8.87	0.81	1.61	4.03	83.87

Once the sample schools were selected for each testing mode (OL and PP) and condition (IM and IEM), the student responses pertaining to the schools were pulled from the database. In addition to matching theta and demographic profiles, the equivalency of the two samples was further evaluated using the raw scores and proportion correct value (p-value) descriptive statistics based on the student level data. Once the samples were confirmed equivalent on these measures, the study evaluated item parameters and student performance on the premise that any difference between groups should be the effect of the mode of administration. The testing mode (OL and PP) and condition (IM and IEM) effects were evaluated in terms of IRT model fit, differential item functioning, and the mean of item parameters (a-discrimination, b-difficulty, c-guessing) and person parameter (student theta).

Results

Similar to the REGD approach, the logic of the designs discussed by Way et al. and those applied here revolves around using the equivalence of the samples as a basis for attributing differences in testing results to mode effects. One of the primary questions investigated in this study is the extent to which the IM and the IEM conditions were capable of creating equivalent samples per testing mode and the extent to which adding an additional predictive criterion to the sampling conditions enables us to more readily create equivalent samples. The first form of evidence of the extent to which we were able to create equivalent samples using both the *internal* and the *internal plus external* conditions was described above. As shown in Tables 2 and 3, we were able to successfully apply the design as intended to create pairs of samples per testing mode and theta level, and the pairs of samples created per testing mode and theta interval reflect similar gender and ethnicity distributions. The gender and ethnicity distributions were generally representative of the ethnicity profiles overall at the state level.

Additional evidence reflecting the degree to which the IM and IEM conditions resulted in equivalent groups per testing mode is provided below based on raw score and p-value descriptive statistics for each sample and condition.

Raw Score and Proportion Correct Descriptive Statistics

The raw scores and p-values descriptive statistics described below further confirmed that the samples created for each mode (OL and PP) were equivalent within each condition (IM and IEM). The detailed results follow.

Raw Score Descriptive Statistics

Table 4 shows the year 2 mean raw score and standard deviation (SD) for all students, per gender, and per ethnicity based on the IM and IEM conditions. As shown in Table 4, based on the IM condition, the mean raw score difference between the OL and the PP samples was only 0.05 raw score points for the overall sample. The difference between the OL and the PP samples was also generally small (less than 2 raw score points) in each of the gender and ethnicity subgroups. Based on the IEM condition, the overall raw score difference between the OL and the PP sample groups was slightly larger (2.32 score points) and same was true for the gender and ethnicity subgroups. These differences suggest that both the IM and the IEM conditions were able to produce nearly equivalent samples per testing mode, but the results were generally closer between the modes based on the IM condition.

Table 4. Raw Score Descriptive Statistics Across Modes and Conditions for Overall and by Subgroup

		Internal Matched (IM)		Internal Plus External Matched (IEM)		Difference	
		S1 (OL)	S2 (PP)	S3 (OL)	S4 (PP)	S1-S2 (OL-PP)	S3-S4 (OL-PP)
Overall	Mean	43.66	43.61	44.85	42.53	0.05	2.32
	SD	12.87	13.84	12.72	14.13	-0.97	-1.41
	N	5940	6015				
Male	Mean	45.08	44.68	46.42	43.74	0.40	2.68
	SD	13.20	14.51	12.95	14.79	-1.31	-1.84
	N	2981	2934	2976	3353		
Female	Mean	42.24	42.62	43.29	41.36	-0.37	1.93
	SD	12.36	13.08	12.29	13.3	-0.72	-1.02
	N	2953	3031	2983	3323		
American Indian	Mean	42.16	42.06	42.6	41.5	0.10	1.10
	SD	12.67	13.15	12.76	13.07	-0.48	-0.31
	N	1228	1178	1246	1146		
Asian	Mean	47.51	45.57	48.91	43.82	1.94	5.09
	SD	13.19	14.62	13.05	15.5	-1.43	-2.45
	N	131	160	129	175		
Black	Mean	39.44	37.92	39.92	35.39	1.52	4.52
	SD	11.96	13.20	11.85	13.32	-1.24	-1.46
	N	663	688	610	995		
Hispanic	Mean	39.13	40.10	42.33	37.25	-0.97	5.08
	SD	12.7	13.94	12.21	13.41	-1.24	-1.2
	N	318	262	290	425		
White	Mean	45.34	45.61	46.63	45.58	-0.27	1.06
	SD	12.74	13.73	12.45	13.73	-0.99	-1.29
	N	3492	3410	3585	3613		

Proportion Correct Descriptive Statistics

Table 5 shows the extent to which the comparison groups created based on the IM and the IEM condition have similar p-values; that is, the extent to which the OL and the PP samples answered similar proportions of the test items correctly. The results showed that, in the IM condition, the mean p-value for the OL and PP samples (S1 and S2) was the same. In the IEM condition, the mean p-value differences for the samples S3 and S4 were slightly larger (.04). This trend is consistent with the mean raw score differences observed above. Like the raw score results, the results here indicated that the IM condition and the IEM condition provided similar results for the OL and PP sample groups, but there were slightly larger differences between the OL and PP sample groups based on the IEM condition.

Table 5. Proportion Correct Descriptive Statistics of Across Samples

Descriptive Statistics	Internal Matched (IM)		Internal Plus External Matched (IEM)		Difference	
	S1 (OL)	S2 (PP)	S3 (OL)	S4 (PP)	IM S1-S2 (OL-PP)	IEM S3-S4 (OL-PP)
Mean	0.61	0.61	0.63	0.59	0.00	0.04
SD	0.15	0.13	0.15	0.13	0.02	0.02
Min	0.30	0.34	0.33	0.32	-0.04	0.01
Max	0.88	0.85	0.89	0.84	0.03	0.05

Calibration Results

The results from Tables 2 through 5 suggest that the OL and PP samples based on the IM condition were equivalent; samples based on the IEM condition also showed fairly equivalent results, though to a lesser extent. As noted, the logic of the design considered in this study indicates that, based on the notion that we have equivalent samples of students to compare across testing modes, any differences observed in the testing outcomes between the OL and PP groups should be an indication of mode effects. Having provided evidence for the equivalence of samples, we now move to discuss evidence of mode effects.

In order to assess mode effects, we calibrated the OL and PP samples separately using the 3PL IRT model and then examined any differences in terms of model fit, differential item functioning (DIF), mean item parameters, and mean theta for the overall sample as well as mean theta per gender and ethnicity subgroups. The results are presented below.

IRT Model Fit and DIF

After the item analyses, the four data sets were calibrated separately and item and person parameters were estimated. Yen's Q statistic (Yen, 1981) was used for flagging items with poor IRT model fit and the Linn and Harnisch (1981) method was used for flagging items for differential item functioning (DIF). A larger number of poor fit items or items flagged for DIF for a given mode may indicate that the items were sensitive to the mode of administration. The results are summarized in subsequent sections.

The calibration results showed no items flagged for poor model fit in sample S1 (OL in IM) and only one item (item 59) was flagged in each sample S2 (PP in IM), S3 (OL in IEM), and S4 (PP in IEM), indicating that, irrespective of the mode of administration, the student responses fit the 3PL model well.

Table 6 presents the number of items flagged for DIF. Only the focal groups that were flagged for DIF are shown in the table. No items were flagged for DIF for gender, for White students, or for American Indian students. Therefore, those subgroups are not presented in the table. Under the IM sampling condition, most of the items that were flagged in the OL mode were also flagged in the PP mode, and in the same direction, except for item 8 for Asian students and item 56 for Black students. Item 8 favored Asian students in the OL mode and item 56 favored Black students in the PP mode. These items, however, did not disfavor the subgroups in the respective alternate modes.

In the IEM sampling condition, some items flagged in the OL testing mode were not flagged in the PP testing mode, and vice versa. Since only four items (items 5, 19, 32, and 45) out of 72 were flagged, the proportion of items flagged is far lower than the chance level. The two items (items 32 and 45) that favored Asian and Black students in the PP mode did not disfavor them in the OL mode. Similarly, item 5 favored the Hispanic student subgroup irrespective of mode. These results suggest that most of the items in the test were not sensitive to the mode of administration.

Table 6. Number of Items Flagged for DIF Across Modes and Conditions

Ethnicity	Internal Matched (IM)		Internal Plus External Matched (IEM)	
	S1 (OL)	S2 (PP)	S3 (OL)	S4 (PP)
Asian	8+, 24-	24-	19-	32+
Black	45+	45+, 56+		45+
Hispanic	5+	5+	5+	5+

Item Parameters

Table 7 shows the average item parameters by testing mode based on the IM and IEM conditions. The average a-parameter reflects the overall discriminating power of the test and the average b-parameter shows the overall difficulty. A smaller average a- or b-parameter suggests that the resulting test was less discriminating (a) or easier (b). Conversely, a larger average a- or b-parameter suggests the resulting test was more discriminating (a) or more difficult (b). Based on both the IM condition and the IEM condition, the PP mode resulted in a more discriminating test: a=0.81 for S1 (OL) versus 0.84 for S2 (PP), 0.80 for S3 (OL) versus 0.89 for S4 (PP). On the other hand, the OL test in the IM condition resulted in a higher average location parameter (-.19) than the PP test (-.21), while in the IEM condition the opposite held; there the average location parameter was higher for the PP group (-.08 versus -.37). However, the difference in the mean b-parameter between the two modes was smaller in the IM condition than in the IEM condition.

Table 7 also shows the root mean square deviation (RMSD) for the IRT parameters between the modes in each condition. The RMSD here measures the average difference between the item parameters in the two modes and is given by:

$$RMSD = \sqrt{\frac{\sum (O\hat{L} - P\hat{P})^2}{n}}$$

where $\hat{OL}$ is the estimated item parameter (b, for example) in the OL mode, $\hat{PP}$ is the estimated parameter in PP mode, and n is the number of items in the test. A larger RMSD may indicate that the parameters estimates were affected by the mode of administration.

The results indicate that the RMSDs for the a- and b-parameter in the IEM condition are almost double those in the IM condition. For example, the RMSDs for a- and b-parameter are 0.10 and 0.23, respectively, in the IM condition and 0.25 and 0.43, respectively, in the IEM condition. Two key conclusions bearing on the IM condition as compared to the IEM condition can be drawn from these results. First, despite the similarity in the raw scores and p-values between the samples, there are some mode effects in the EOI test, as indicated by the differences in the a- and b-parameter estimates in the IM condition and the IEM condition. Second, the larger RMSD in the IEM condition may suggest that the addition of an external criterion for the purposes of creating equivalent samples did not enhance our ability to produce equivalent samples; it rather detracted.

Table 7. Average IRT Parameters and RMSD of Item Parameters Between Mode

Item Parameters	Internal Matched (IM)			Internal Plus External Matched (IEM)		
	S1 (OL)	S2 (PP)	RMSD	S3 (OL)	S4 (PP)	RMSD
A	0.81	0.84	0.10	0.80	0.89	0.25
B	-0.19	-0.21	0.23	-0.37	-0.08	0.43
C	0.21	0.21		0.21	0.21	

Student Performance

Table 8 lists the mean theta in the 0/1 metric based on the IM condition and the IEM condition. The table also lists the difference of the mean theta between testing modes for the two conditions and effect sizes (ES). The effect size statistic measures the standardized mean difference of thetas between the OL and PP sample groups in the two conditions. The effect size was calculated as:

$$ES = \frac{\overline{OL} - \overline{PP}}{\sqrt{\dfrac{s_{OL}^2(n_{OL}-1)+s_{PP}^2(n_{PP}-1)}{n_{OL}+n_{PP-2}}}}$$

where $\overline{OL}$ is mean theta of OL mode, $\overline{PP}$ is mean theta of PP mode, s_{OL}^2 and n_{OL} are variance and the sample size for OL mode and s_{PP}^2 and n_{PP} are variance and the sample size for PP mode, respectively. A larger effect size indicates that the mean theta difference between the OL and PP samples was large.

Table 8. Mean Theta Across Modes and Conditions

		Internal Matched		Internal Plus External Matched		Difference and Effect Size			
		S1 (OL)	S2 (PP)	S3 (OL)	S4 (PP)	S1-S2	ES	S3-S4	ES
Overall	Mean	-0.20	-0.29	-0.26	-0.28	0.09	0.08	0.02	0.02
	SD	1.17	1.21	1.19	1.22	-0.04		-0.04	
	Alpha	0.92	0.93	0.92	0.93				
	N	5940	6015	5961	6726				
Male	Mean	-0.09	-0.21	-0.12	-0.2	0.12	0.09	0.07	0.06
	SD	1.23	1.30	1.24	1.31	-0.07		-0.07	
	N	2981	2934	2976	3353				
Female	Mean	-0.32	-0.36	-0.40	-0.35	0.04	0.04	-0.05	-0.04
	SD	1.10	1.12	1.12	1.12	-0.02		-0.01	
	N	2953	3031	2983	3323				
American Indian	Mean	-0.34	-0.40	-0.47	-0.34	0.06	0.05	-0.13	-0.11
	SD	1.15	1.12	1.18	1.10	0.03		0.09	
	N	1228	1178	1246	1146				
Asian	Mean	0.10	-0.16	0.10	-0.22	0.26	0.20	0.33	0.23
	SD	1.28	1.32	1.28	1.42	-0.04		-0.14	
	N	131	160	129	175				
Black	Mean	-0.58	-0.80	-0.73	-0.91	0.22	0.19	0.18	0.15
	SD	1.11	1.24	1.13	1.27	-0.14		-0.14	
	N	663	688	610	995				
Hispanic	Mean	-0.64	-0.61	-0.51	-0.70	-0.03	-0.02	0.19	0.17
	SD	1.18	1.27	1.10	1.17	-0.10		-0.07	
	N	318	262	290	425				
White	Mean	-0.05	-0.11	-0.09	-0.02	0.06	0.05	-0.08	-0.06
	SD	1.15	1.19	1.16	1.15	-0.03		0	
	N	3492	3410	3585	3613				

Overall, the mean theta values in the IM condition differ by about .09 and in the IEM condition by .02 with effect size of .08 and .02 respectively. The effect size of .08 for the mean theta between the two modes in the IM condition can be interpreted as indicating that the average student in the OL sample exceeded the theta score in the PP sample by about 8%. The mean theta for the OL sample was higher than the PP sample for the overall student group and across most of the subgroups in both the IM and IEM conditions, indicating that the OL samples performed relatively higher than the PP samples. The largest effect size for the mean theta was observed for Asian and Black student subgroups both in favor of the OL samples meaning that these student subgroups performed higher in OL mode.

When the overall difference is translated into the scale score metric with the slope and intercept of the linear equation set at 50 and 500, the difference for all students overall is about 4 scale score points in the IM condition and only 1 scale score point in the IEM condition. These differences can be considered small.

Summary and Discussions

Online and paper and pencil comparability studies using non-random sampling designs are recent developments. Researchers have suggested use of internally and externally matched samples in situations where the prior student level performance information is available. However, no literature is available for the situation we examined here, where only *school-level* prior academic and demographic information were available, on an end-of-instruction test.

The current study created equivalent samples to compare mode effects by matching school-level theta scores and demographic profiles between two administrations. Two designs were reviewed: the internal matching approach using scores on a given content area to predict scores from the same content area, and the internal plus external matching approach where an additional content area was used to predict scores in a subsequent year. The study compared the extent to which these designs yielded equivalent groups per testing mode and evaluated mode effects. The equivalence of the samples was evaluated primarily in terms of raw score and p-value descriptive statistics. Mode effects were analyzed through IRT model fit, DIF, and item and person parameters. The examination of item and person parameters also included RMSD and standardized mean difference (effect size).

The raw score and p-value descriptive statistics provided evidence of equivalence in the OL and PP samples in the IM condition. Results based on the IEM condition were also reasonably equivalent, though not as much as in the IM condition.

The results further showed that despite the similarity of the raw score and p-value descriptive statistics between the two modes, the resulting IRT item parameters were slightly different between the two modes. The difference in parameters between the OL and PP samples was larger in the IEM condition than in the IM condition, as shown by the mean a- and b-parameters in the samples and the corresponding RMSD values. As indicated, the RMSD was almost double for the IEM condition compared to the IM condition. In summary, these results suggest that matching the prior performance and demographic information of the same content (internal matching) to create equivalent samples may be sufficient, and the addition of the external criteria may not add significant information.

In terms of mode effects, the item parameter and mean theta results showed some differences between the samples. However, the apparent effect size was small, and in summary, the test results did not show statistically discernible mode effects based on model fit, DIF, or student performance, despite some differences in item parameters.

The evidence provided on IM and IEM methods suggests that where psychometricians and state assessment officials are looking for ways to create equivalent samples to compare across testing modes, without the need to do so through randomization and while still obtaining large sample sizes, the IM method may be sufficient, and including an additional external criterion, and accordingly adding costs, did not appear to provide additional benefits.

As noted, comparability studies using non-random sampling designs are a recent development. While we have provided some additional evidence bearing on one of the more interesting approaches developed in recent years, study in this area should continue. In order to gather additional insight into the practical value or potential limits of the IM approach, it may be useful to consider designs that afford an opportunity to compare the IM approach to other designs such as REGD. This kind of analysis may enable researchers to evaluate how the equivalent samples created under the two approaches compare; whether a design such as REGD, which, as noted, can be costly, provides greater precision in creating equivalent groups; and the extent to which the results under an IM design differ. This kind of research could clarify relative value of the options we have when building a foundation for detecting mode effects in comparability studies.

References

Ito, K., & Sykes, R. C. (2004). Comparability of Scores from Norm-Referenced Paper and Pencil and Web-Based Linear Tests for Grades 4 – 12. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Keng, L., McClarty, K. L., & Davis, L. L. (2006). Item-Level Comparative Analysis 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.

Linn, R. L. & Harnisch, D. L. (1981). Interactions between item content and group membership and achievement test items. Journal of Educational Measurement, 18(2), 109-118.

Mead, A.D. & Drasgow, F. (1993). Equivalence of computerized and paper cognitive ability tests: A meta-analysis. Psychological Bulletin, 114(3), 449-458.

Peak, P. (2005). Research trends in comparability studies. Pearson Educational Measurement. http://www.pearsonedmeasurement.com/research/research.htm

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. Journal of Technology, Learning, and Assessment,3(6). Available from http://www.jtla.org.

Wang, S. (2004). Online or Paper: Does Delivery Affect Results? Administration Mode Comparability Study for Stanford Diagnostic Reading and Mathematics Tests. Harcourt Assessment Report.

Way, W. D., Davis, L. L., Fitzpatrick, S. (2006). Score Comparability of Online and Paper Administrations of the Texas Assessment of Knowledge and Skills. PEM Research report 06-01, April 2006.

Way, W. D., Um, K., Lin, C., & McClarty, K. L. (2007, April). *An Evaluation of a Matched Samples Method for Assessing the Comparability of Online and Paper Test Performance*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Yen, W.M. (1981) Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245-262.