Does Matching Quality Matter in Mode Comparison Studies?

Educational and Psychological

Measurement

2015, Vol. 75(6) 1045–1062

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414565006

epm.sagepub.com

Ji Zeng¹, Ping Yin², and Kerby A. Shedden³

Abstract

This article provides a brief overview and comparison of three matching approaches in forming comparable groups for a study comparing test administration modes (i.e., computer-based tests [CBT] and paper-and-pencil tests [PPT]): (a) a propensity score matching approach proposed in this article, (b) the propensity score matching approach used by Lottridge, Nicewander, and Mitzel, and (c) a modified approach of matched samples comparability analyses (MSCA) mentioned by Way, Davis, and Fitzpatrick. Different matching approaches resulted in different matched data with differing degrees of matching quality, and matched data from each matching approach were then used in the mode comparison investigation. Construct equivalence was examined and the level of invariance was found to be consistent across modes for all three matching approaches. Raw-to-scale score conversion tables were created, and the impact on CBT students' proficiency classification was examined. The comparison of the number of CBT students whose proficiency classification would be affected and the equality of score distributions between modes on raw scores and scale scores across the three matching approaches indicate that the propensity score matching approach delineated in this article led to the most consistent evidence for the conclusion of the mode comparison.

Keywords

propensity score matching, mode comparison, conversion tables

¹Michigan Department of Education, Lansing, MI, USA

²Questar Assessment Inc., Apple Valley, MN, USA

³University of Michigan, Ann Arbor, MI, USA

Corresponding Author:

Ji Zeng, Division of Accountability Services, Michigan Department of Education, P.O. Box 30008, Lansing, MI 48909, USA.

Email: zengj@michigan.gov

There is an increasing trend and desire to move statewide assessments from paper-and-pencil tests (PPT) to computer-based tests (CBT). For example, several consortia have announced plans to implement CBT, including new general assessments in English Language Arts and Mathematics that are currently being developed by both the Partnership for Assessment of Readiness for College and Careers and the Smarter Balanced Assessment Consortium; new alternative assessments under development by the Dynamic Learning Maps Alternate Assessment System Consortium and the National Center and State Collaborative; and new English language proficiency tests under development by the World-Class Instructional Design and Assessment consortium and the English Language Proficiency Assessment for the 21st Century consortium. In addition, states not participating in the consortia area also moving their assessments online (such as Utah, Texas, and New York).

Despite plans to implement CBT assessments in the 2014-2015 or the 2015-2016 academic years, PPT and CBT tests are likely to coexist for some time for some states because not every school can meet the technology and infrastructure requirements for those CBT assessments (Way, Lin, & Kong, 2008). According to the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), score comparability needs to be established whenever a test is administered in different modes. In practice, one critical decision to be made is whether data from the two different modes can be combined to create a single set of item parameters. When the Rasch or one-parameter logistic model (1PL) and/or the Partial Credit Model (PCM) are used, a single set of item parameters leads to a single raw-to-scale score conversion table for reporting. The decision on whether it is appropriate to combine data for a single calibration is important from the practical point of view (due to its impact on students and schools under accountability systems), but is underrepresented in past research on mode comparison.

The remainder of the introduction includes (a) a brief review of the previous mode comparison literature and (b) a brief overview of the three matching approaches under consideration in this article. The review of mode comparison literature is structured as follows. First, we summarize the analytical methods used in past research for judging mode comparability, and we argue that mode comparison in this study should focus on test-level comparability. Second, we review comparison study designs, with a focus on how comparable samples are formed. We then introduce the three different matching approaches considered in this article for forming comparable samples.

Methodologies for Comparability Analysis

Most mode comparison studies concern test-level (overall score) comparability (e.g., Lottridge, Nicewander, & Mitzel, 2011; Randall, Sireci, Li, & Kaira, 2012; Schroeders & Wilhelm, 2011), but some also investigate item-level comparability (e.g., Keng, McClarty, & Davis, 2008; Randall et al., 2012), or content domain-level comparability (e.g., Kim & Huynh, 2007, 2008). Some studies have also examined

different aspects of tests to establish comparability across modes, such as the comparison of cross-mode correlations (e.g., Lottridge et al., 2011; Neuman & Baydoun, 1998; Schroeders & Wilhelm, 2011), mean scale scores (e.g., Bennett et al., 2008; Kim & Huynh, 2007, 2008; Lottridge et al., 2011), cumulative distributions (e.g., Puhane, Boughton, & Kim, 2007), test characteristic curves and test information functions (e.g., Bennett et al., 2008; Kim & Huynh, 2007, 2008), reliabilities (Lottridge et al., 2011; Neuman & Baydoun, 1998; Schroeders & Wilhelm, 2011), and construct equivalence (e.g., Kim & Huynh, 2007, 2008; Lottridge et al., 2011; Neuman & Baydoun, 1998; Randall et al., 2012; Schroeders & Wilhelm, 2011).

As this study focuses on whether matching quality matters in mode comparison investigations in terms of the impact on students' proficiency classification consistency and reported scale scores, we investigated only construct equivalence, raw-to-scale score conversion tables, and cumulative distributions on raw scores and scale scores. Therefore, we examined only the overall test-level comparability.

If comparability at the test level can be established, CBT students are then combined with PPT students for a single run of item calibration and for creating a single raw-to-scale score conversion table (when Rasch/IPL and/or PCM are used for calibration). If comparability at the test level cannot be established, differences across modes are usually not constant across the ability scale (e.g., as shown in test characteristic curves and test information functions in Kim & Huynh, 2007, 2008). Consequently, separate item calibrations are conducted and separate conversion tables with one for each mode are created for score reporting.

Comparability Study Designs

Most mode comparability studies at the K-12 level are done using a single group with counterbalancing design (Lottridge et al., 2011). This design tries to single out the mode difference for a possible causal inference on the mode effect. However, such a design is very difficult to implement in practice for state assessments (Kingston, 2009). Consequently, some statistical methods have been proposed to obtain comparable PPT and CBT samples without the need for students to take a test in both modes. Two approaches are matched samples comparability analyses (MSCA; Way, Davis, & Fitzpatrick, 2006; Way et al., 2008) and propensity score matching (e.g., Lottridge et al., 2011). Propensity score matching and the MSCA method are similar in attempting to create statistically matched samples (Way et al., 2008), though the latter seems to match only on previous achievement scores (Way et al., 2006, 2008). A modified MSCA method was considered in this study, as well as the propensity score matching approach used by Lottridge et al. (2011). Details of these two approaches and a different propensity score matching approach delineated here are provided below. The last approach is introduced first and is described within the framework of what needs to be considered in conducting a propensity score matching.

The Optimal Pair Matching With Omnibus Balance Test (OPM-OBT) Approach. This approach tries to match students using propensity scores. A propensity score, which does not depend on information regarding the outcome of interest, is the conditional probability of assignment to treatment (in our case, taking CBT instead of PPT) given various covariates (Rosenbaum & Rubin, 1985). With a dichotomous treatment variable, logistic regressions with the treatment assignment as an outcome are used to estimate propensity scores (e.g., Harder, Stuart, & Anthony, 2010). In this article, CBT was coded as 1 and PPT was coded as 0. Various approaches such as matching, weighting, and subclassification can be applied to form comparable groups (Harder et al., 2010; Stuart, 2010) after the propensity scores are estimated. We only considered matching, and in particular, pair matching in forming comparable groups with the same sample size to eliminate possible effects of different sample sizes on item calibration in the item response theory (IRT) framework.

Five issues need to be considered when conducting propensity score matching: (a) choice of covariates, (b) dealing with missing data on the covariates, (c) matching methods, and (d) assessing the matching quality (Lottridge, et al., 2011; Steiner, Cook, Shadish, & Clark, 2010; Stuart, 2010). In addition, unlike randomization which can balance both observed and unobserved covariates, matching on propensity scores can only balance observed covariates (Rosenbaum & Rubin, 1985). Therefore, the fifth issue to consider when conducting propensity score matching is (e) the possible violation of ignorable treatment assignment after analyzing the outcome of interest.

Choice of covariates. To fully use the capability of propensity scores in balancing multiple covariates, we included the following in our propensity score estimation model: all possible recent previous achievement scores, all available demographic variables at the student level, and the school-level background variables such as the number of students,¹ the number of computers with high-speed Internet per student, and school-level achievement variables (created as the mean of the school-level aggregates of the previous 3 years) for the targeted students (i.e., ELLs) only. In addition, quadratic terms of continuous covariates and interactions between student demographic variables and student achievement variables were also considered in the model. Note that since different general assessments are administered at different grade levels, the number of previous achievement variables included in the model for each grade level is different.

When comparing students nested within schools, a more appropriate choice seems to be using a multilevel design (e.g., Hong & Raudenbush, 2006). However, with our data, more than half of the schools had very few (≤ 5) ELLs, so a multilevel method was not practical for the data under investigation. To still take school-level covariates into consideration in estimating the propensity scores, the school-level covariates were modeled as individual-level covariates.

Dealing with missing data on the covariates. Since very few students took CBT in our data and about one third of them were found to have missing data on some previous achievement variables, we used imputation for all students (as PPT students also

have relatively high rates of previous achievement missing). This was done to (a) keep all CBT students in the analyses and (b) to keep the full PPT pool for matching.

In our study, a single imputation was carried out using the R package MICE (van Buuren & Groothuis-Oudshoorn, 2011), which conducts multivariate imputation by chained equations. However, instead of using the program's capability of creating multiple sets of imputed values, we only used one set of imputed values, as the comparison between multiple imputations and single imputation is out of the scope of this article. Since there are both school-level and student-level covariates, we conducted imputations in two steps. First, imputations were conducted at the school level. Then the imputed school-level data were merged with student-level variables to conduct imputations at the student level.

All matching approaches considered here used the same imputed data from this step. Therefore, the differences in matching results are purely the result of the differences in matching approaches themselves.

Matching methods. Different matching methods exist in the literature, such as the nearest neighbor procedure used in Lottridge et al. (2011). The optimal matching algorithm, which "optimize global, rather than local, objectives" (Hansen, 2004, p. 612), is found to often perform better than the nearest neighbor procedure for pair matching with a large pool of controls (Hansen, 2004). Since we had an extensive pool of PPT students in comparison with CBT students in our data, the optimal matching algorithm was implemented using the R package OPTMATCH (Hansen & Klopfer, 2006).

Assessing the matching quality. Checking the quality of the resulting matched samples is the most important step in using matching methods (Stuart, 2010). Existing methods include, but are not limited to, the computation of standardized bias and two-sample *t* tests before and after matching (Caliendo & Kopeinig, 2008). For the optimal matching method chosen in this article, the balance check approach used in the R package Rltools (Bowers, Fredrickson, & Hansen, 2010) was used because of its ability to test balance not only on each individual covariate, but also on all linear combinations of the covariates in the propensity score model (Hansen & Bowers, 2008). Balance tests that show statistically significant results indicate a need to rebuild the propensity score estimation model with different covariates or different functional forms of the covariates.

The Lottridge et al. (2011) Approach. The propensity score matching used by Lottridge et al. (2011) chose covariates (with the majority being individual-level variables, and one variable being a district-level variable) by using a forward stepwise logistic regression procedure with the inclusion threshold being set at chi-square probability of .05. After obtaining the estimated propensity scores from the final logistic regression model, the nearest neighbor procedure was used to form a matched sample of PPT students. In essence, the nearest neighbor procedure identifies a PPT student who has the closest estimated propensity score with a specific CBT student under investigation, and once such a PPT student is identified, he or she will be removed

from the PPT pool for further matching consideration (Lottridge et al., 2011). We employed this approach strictly following these descriptions, except that the R package MatchIt (Ho, Imai, King, & Stuart, 2011) was used to carry out the nearest neighbor matching algorithm.

A Modified MSCA. The MSCA approach described in Way et al. (2006) involves matching on identical scores from the previous year's achievement test(s). A little explanation is in order here. The assessment being investigated is the English Language Proficiency Assessment, or ELPA, for ELLs. Available covariates include general assessments administered to all students in mathematics, reading, writing, science, and social studies. Since we have different general assessments (i.e., math and reading administered to Grades 3 to 8 and Grade 11, writing administered only to Grades 4, 7, and 11, science administered only to Grades 5, 8, and 11, and social studies administered only to Grades 6, 9, and 11) at different grade levels, when implementing this approach, we tried to match exactly on the previous ELPA achievement, while simultaneously satisfying the condition that the differences on math and reading prior achievement be minimal.² To achieve exact *N* count match from the PPT to CBT groups at each grade level, we first used a stratified random sampling approach (with each previous year's ELPA scale score point functioning as a stratum). Within each stratum, for students with prior achievement information in both math and reading general assessment, we computed a distance measure for each PPT student using the following distance formula:

$$\text{Distance} = (\text{PPT_Reading} - \text{CBT_Reading})^2 + (\text{PPT_Math} - \text{CBT_Math})^2 \quad (1)$$

Within each ELPA scale score stratum, the PPT student with the smallest distance value was chosen as the match for the given CBT student.

Purpose of This Study

Among these three matching approaches, the OPM-OBT approach (the approach proposed in this article) is the most time consuming, while the modified MCSA is the least time consuming. Since in the near future, multiple mode comparison studies are expected to be carried out in a short amount of time to make sure that score reports can be delivered on time if not much earlier; we attempted to investigate whether matching quality matters in a mode comparison study by addressing the following two research questions:

Research Questions 1: How different is the matching quality resulting from the three matching approaches?

Research Questions 2: If the matching quality differs dramatically, do such differences still lead to similar mode comparison results or do they lead to considerably different mode comparison results?

If differing degrees of matching quality still lead to similar mode comparison conclusions, the least time consuming method would be recommended. However, if mode comparison results are different, the approach leading to the best matching quality would be recommended.

Method

Data

To investigate the impact of matching quality on mode comparison results, data from a large-scale state ELPA testing program from Grades 3 to 12 was used. On the ELPA, operational items (for score reporting) are exactly the same for each of the following grade spans: 3 to 5, 6 to 8, and 9 to 12. The testing program was also designed to have a vertical scale from Kindergarten to Grade 12. Within each grade span, data from all grade levels were calibrated together using the Rasch/1PL and the PCM. In addition, a fixed parameter approach for anchoring items repeated from a previous administration was used as the equating method. As mentioned above, when the Rasch/1PL and/or the PCM are used for IRT calibrations, raw-to-scale score conversion tables can be made for score reporting.

In spring 2011, a CBT version of this testing program for Grades 3 to 12 was introduced with a voluntary participation at the school level. Since performance from the CBT participation was used for score reporting, an operational decision had to be made regarding whether these students' performance data (i.e., CBT data) could be combined with other students' performance data (i.e., PPT data) to create only one conversion table, or whether the different test delivery modes have a sufficiently strong impact on student performance that separate conversion tables are more appropriate for score reporting.

Table 1 presents information on the number of students taking CBT and PPT from Grades 3 to 12, showing that across all these grade levels, approximately 2% to 5% of students at each grade level took the CBT.

Analysis Procedures

After obtaining matched data from the three different matching approaches, two sets of analyses were conducted: (a) matching quality comparison and (b) mode comparison. Matching quality was compared by using a *t* test on continuous covariates and the chi-square test on dichotomous covariates between modes before and after each matching approach was applied. The smallest number of unbalanced covariates after matching indicates the best matching quality, while the largest number of unbalanced covariates after matching indicates the worst matching quality.

The second set of analyses investigated whether differing degrees of matching quality lead to different conclusions in mode comparison studies (i.e., whether one conversion table using combined data is appropriate). Three analyses were conducted for this purpose.

Table 1. CBT and PPT Participation Information From Grades 3 to 12.

Grade	Number of CBT students		Number of PPT students	
	<i>n</i>	%	<i>n</i>	%
3	308	4.5	6,603	95.5
4	243	4.2	5,574	95.8
5	271	5.4	4,772	94.6
6	144	3.6	3,902	96.4
7	121	3.1	3,721	96.9
8	136	3.7	3,559	96.3
9	196	4.7	4,010	95.3
10	152	4.1	3,532	95.9
11	139	5.1	2,593	94.9
12	115	5.3	2,072	94.7

Note. CBT = computer-based test; PPT = paper-and-pencil test.

First, nested models were compared with each other to establish construct equivalence (Schroeders & Wilhelm, 2011): configural invariance, strong invariance, and strict invariance. Similar to Schroeders and Wilhelm (2011), here we used Mplus 7.11 (Muthén & Muthén, 2012) to estimate all models using the default estimator—weighted least squares mean and variance adjusted estimator with Theta parameterization. Because of the problems found with the chi-square statistics (Chen, 2007; Cheung & Rensvold, 2002), the following fit indices and cutoff criteria were used: the comparative fit index (CFI) $\geq .95$ and the root mean square error of approximation $< .05$ (Hu & Bentler, 1998) for indicating a good model fit; and a change of $\geq -.010$ in CFI (Chen, 2007; Cheung & Rensvold, 2002) for indicating invariance for each step of the nested model comparison. Since our main concern is the comparability of CBT and PPT, we focused on whether CBT and PPT displayed the same level of invariance.

Next, the Rasch/1PL and the PCM IRT calibration were carried out using WINSTEPS 3.68.2 (Linacre, 2009) and the resulting raw-to-scale score conversion tables for the two groups (i.e., CBT vs. each matched PPT group) were compared. Student achievement and growth scores on the ELPA were used for accountability purposes. We thus did related analyses focusing on how student status and growth may be affected by different conclusions on mode comparison. Because a large proportion of students took PPT, no practical differences were found on reported scale scores for PPT students. Therefore, the growth computation, in practical sense, can only differ for CBT students, and can only happen when an operational decision was made to apply a separately calibrated conversion tables to them. This, however, has nothing to do with the matching quality comparison, unless differing degrees of matching quality lead to different mode comparison conclusions. It then follows that the proficiency status is the focus here. For proficiency status, two investigations were carried out:

1. Based on the various conversion tables created, the number of CBT students classified differently using the CBT conversion table and each of the matched PPT conversion tables was compared
2. The distributions of raw scores and scale scores from CBT and each matched PPT set were compared

Finally, consistency between results from the raw score distribution equality test and the scale score distribution equality test was examined for each of the three matching approaches.

Because the Rasch/1PL and the PCM were used, a one-to-one relationship was established between raw score points and scale scores. In other words, it is expected that if the cumulative distributions of raw scores are similar between the two modes, the scale scores should be similar as well. Therefore, the Kolmogorov–Smirnov test for two independent samples was chosen for testing the equality of distributions from the two modes. The null hypothesis and the alternative hypothesis are (Sheskin, 2011, p.596):

$$\begin{aligned} H_0: F_1(X) &= F_2(X) \text{ for all values of } X, \text{ and} \\ H_1: F_1(X) &\neq F_2(X) \text{ for at least one value of } X. \end{aligned} \quad (2)$$

The statistics being tested here is the vertical distance between the two cumulative probability distributions of the two groups, with the alternative hypothesis being that the largest vertical distance is larger than what would be expected by chance, if the two groups are random samples from the same population.

Results

The purpose of this article is twofold. First, we compared the performance of the three matching approaches on their ability to effectively balance covariates across modes. Second, we investigated how such differences in matching quality impact mode comparison results, and in particular, how differences in reporting scales influence students' proficiency classifications and how scale scores may be different (and thus may affect their growth computation).

Table 2 presents the results of the balance check on all covariates, and in particular, the number of covariates found to be statistically significant at $\alpha = .05$ level before and after matching. According to Table 2, the OPM-OBT approach performed the best, as it managed to balance all covariates at all grade levels except two covariates at Grade 12. On the contrary, the modified MSCA performed the worst, and sometimes even led to a larger number of unbalanced covariates in comparison to before matching. The propensity score matching approach used by Lottridge et al. (2011) performed quite well and led to few unbalanced covariates after matching.

The results from the construct equivalence investigation indicated that the hypothesis of CBT and PPT sharing the same constructs for operational items was not rejected at all grade spans and for all matched data resulting from the three different

Table 2. Balance Check of Covariates From Grades 3 to 12.

Grade	Number of significantly different covariates (at $\alpha = .05$)			
	Before matching	After matching		
		Modified MSCA	Lottridge et al. (2011)	OPM-OBT Approach
3	13	12	4	0
4	12	18	4	0
5	11	18	3	0
6	18	23	2	0
7	15	15	4	0
8	13	14	1	0
9	9	6	0	0
10	6	5	2	0
11	18	14	0	0
12	17	16	3	2

Note. MSCA = matched samples comparability analyses; OPM-OBT= optimal pair matching with omnibus balance test.

matching approaches. Therefore, the differing degrees of matching quality reported above was not found to be associated with different findings with regard to construct equivalence. Since this is a required check but not the focus in this article, detailed results are not reported.

Table 3 presents differences in raw-to-scale score conversions at the proficiency cut at each grade level. Consistent with what can be expected from the covariates balance results, the two propensity score matching approaches (i.e., the Lottridge et al., 2011, and the OPM-OBT approach) are most alike in classifying students based on the raw scores associated with the proficiency cuts. However, we also observed that the scale scores at the proficiency cuts are most alike for the modified MSCA and the Lottridge et al. (2011) approach. We thus inferred that a different nonlinear relationship was established between raw scores and scale scores for the three matched PPT data sets.

As mentioned above, with our data, we found no practical differences on reported scale scores for PPT students when all of them were considered (i.e., they were not affected when they were calibrated by themselves or calibrated together with CBT students). Therefore, proficiency classification can only differ for CBT students when a separate conversion table is used for them. To make a fair comparison for the voluntarily participating CBT group, we investigated the possible raw-to-scale score conversion relations established by different matched PPT groups rather than using all PPT students (i.e., without applying any matching approaches). We concluded from the number of CBT students who could have a classification change if a matched PPT conversion was applied to them (as shown in Table 4) that the Lottridge et al. (2011) approach led to the largest number of CBT students with a proficiency

Table 3. Raw-to-Scale Score Conversions at Proficiency Cuts From Grades 3 to 12.

Grade	Proficiency cut on scale score	Scale scores at or above proficiency cuts				Corresponding raw scores			
		CBT	Modified MSCA	Lottridge	OPM-OBT	CBT	Modified MSCA	Lottridge	OPM-OBT
3	619	620	619	620	620	57	57	58	58
4	626	627	627	627	627	60	60	61	61
5	633	633	635	633	633	62	63	63	63
6	635	636	637	637	635	64	65	65	64
7	641	642	643	641	641	66	67	66	66
8	648	649	651	650	648	68	69	69	68
9	658	659	658	658	659	66	66	66	66
10	661	663	661	661	662	67	67	67	67
11	664	666	665	664	665	68	68	68	68
12	672	674	672	672	673	70	70	70	70

Note. CBT = computer-based test; MSCA = matched samples comparability analyses; OPM = Optimal Pair Matching; OPM-OBT= optimal pair matching with omnibus balance test. The calibrations and conversion tables were made for each of the following groups: CBT students and each set of matched PPT students resulting from the modified MSCA approach, Lottridge et al. (2011) approach, and the OPM-OBT approach. The number of students involved in each of the four calibrations was exactly the same per grade span, and the number of students per grade for each data set was the same as the number reported for CBT students in Table 1.

Table 4. Number of CBT Students With Proficiency Classification Change After Applying Different Conversion Tables Resulting From Each of the Matched PPT Samples.

Grade	Modified MSCA	Lottridge	OPM-OBT
3	0	11	11
4	0	10	10
5	13	13	13
6	5	5	5
7	6	6	0
8	1	1	1
9	0	0	0
10	0	0	0
11	0	0	0
12	0	0	0

Note. CBT = computer-based test; MSCA = matched samples comparability analyses; OPM-OBT= optimal pair matching with omnibus balance test.

Table 5. Tests of Equality of Distributions of Raw and Scale Scores of CBT and Matched PPT Samples.

Grade	Raw score			Scale score		
	Modified MSCA	Lottridge	OPM-OBT	Modified MSCA	Lottridge	OPM-OBT
3	0.244***	0.075	0.084	0.289***	0.039	0.055
4	0.226***	0.041	0.045	0.259***	0.058	0.066
5	0.181***	0.066	0.074	0.218***	0.066	0.074
6	0.174*	0.104	0.111	0.208**	0.118	0.146
7	0.198*	0.157	0.140	0.223**	0.190*	0.165
8	0.096	0.118	0.059	0.125	0.169*	0.118
9	0.138*	0.061	0.087	0.163*	0.087	0.107
10	0.171*	0.132	0.158*	0.197**	0.158*	0.184*
11	0.079	0.079	0.072	0.079	0.086	0.101
12	0.096	0.087	0.070	0.096	0.113	0.113

Note. CBT = computer-based test; MSCA = matched samples comparability analyses; OPM-OBT= optimal pair matching with omnibus balance test.

*** $p < .001$. ** $p < .01$. * $p < .05$.

classification change, while the modified MSCA approach led to the smallest number of CBT students with a proficiency classification change. Moreover, the performance of the two propensity score matching approaches was found to be similar in this regard.

Table 5 reports the largest vertical distance between the two cumulative probability distributions and the corresponding probabilities of the distribution equality test. Based on Table 5, we could not reject the null hypothesis that raw scores from CBT

and PPT matched groups are random samples from the same population for the Lottridge et al. (2011) approach at any grade level. Similarly, we could not reject this null hypothesis for the OPM-OBT approach at any grade level except Grade 10. However, this null hypothesis was rejected 7 out of 10 times for the modified MSCA approach. Note that conversion tables are made within each grade span rather than at each grade level. Therefore, based on this test, we concluded that modes were not comparable based on matched data resulting from the modified MSCA approach, but were comparable based on data obtained from the two propensity score matching approaches.

When results from the distribution equality test for scale scores are compared with those for raw scores, they are nearly identical for both the modified MSCA approach and the OPM-OBT approach, but were different for the Lottridge et al. (2011) approach. Specifically, the null hypothesis that the two distributions were random samples from the same population was rejected 3 out of 10 times for scale scores, but was rejected 0 out of 10 times for raw scores for the Lottridge et al. (2011) approach. This again confirms that different raw-to-scale score relations were established based on different matched sets.

In conclusion, we found that the OPM-OBT approach performed best in balancing covariates, but the modified MSCA resulted in the smallest number of CBT students being classified differently for proficiency. In addition, we found that various tests show similar results for the two propensity score matching approaches, and there is greater consistency between raw score distributions and scale score distributions across CBT and PPT groups based on matched samples produced by the OPM-OBT approach.

Discussion

Summary

The purpose of this article was to provide a brief overview and comparison of three different matching approaches in forming comparable groups for a test administration mode comparison study. Results suggest that the three approaches gave generally similar conclusions, especially with regard to the construct equivalence across modes and consistency in proficiency classification. However, there were notable differences in their capability to balance covariates and in their performance in establishing relationships between raw scores and scale scores. Based on various analyses, different conclusions could be made using data resulting from the modified MSCA. Specifically, the corresponding proficiency classification consistency would lead to the conclusion that the two modes are comparable, but the distribution equality test (on both the raw score distributions and the scale score distributions) would lead to the conclusion that the two modes are not comparable. The two propensity score matching approaches were found to perform in a similar fashion to each other on all analyses included in this article. However, when consistency of distribution equivalence test of raw scores and scale scores were considered, the Lottridge et al. (2011)

approach was found to have the least consistency. Therefore, the OPM-OBT approach stands out to be the best performing matching approach among the three examined here.

The propensity score matching approaches can easily be used for other comparisons. For example, in the near future, all students may take the CBT, but with different devices. As a result, some students may take the CBT with a virtual keyboard, while others take the CBT with a physical keyboard.

Based on this study, the propensity score matching techniques are in general found to be effective in forming comparable groups, with the approach proposed here performing better overall. However, as pointed out by Rosenbaum and Rubin (1985), propensity score approaches can only balance observed variables. Nevertheless, if unobserved variables are highly correlated with observed variables used for the propensity score estimation, bias of those unobserved variables would be removed somewhat to the extent that they are correlated with the observed ones (Rosenbaum & Rubin, 1984).

Considering Possible Violation of Ignorable Treatment Assignment

This section is included for completeness of carrying out the propensity score matching procedures.

In our study, we tried to include all previous achievement scores, in addition to the scores on the previous year's ELPA. Spring 2011 was the first year the CBT was available, and because of the need for administering a component of the test on a one-on-one basis, only schools with a small number of ELLs and with many computers with high-speed internet were expected to participate. Therefore, we included these two variables in our propensity score estimation model, and we did find such schools tended to participate at an overrepresentative rate. We also included student demographic variables as the ELL population is very diverse. We hoped that including those variables as covariates would help to remove some bias with regard to different computer-person interaction experience due to culture differences. The only possible important variable we did not include (based on comparing with mode comparison literature such as Bennett et al. [2008] and Lottridge et al. [2011]) was student computer skill scores, as this information has not been collected at the state level in the state from which the data came.

Because computers are becoming more widely used in personal life around the world, we did not expect such skills to be excessively different among ELLs (coming from different countries), especially because students were allowed to opt out of taking the CBT and take the PPT instead if they felt more comfortable with PPT. Therefore, all students who took the CBT could be considered as having a high enough comfort level with taking the CBT, and their computer skills should not have harmed their performance.

However, because covariate related to computer skills was not included in the current study and because the literature on the relation between computer skills and

English language ability seems to be sparse, it would have been difficult to conduct a sensitivity analysis with regard to this missing covariate. We believe, however, that since the PPT is the most familiar test delivery mode to students around the world, taking the CBT could only possibly depress student scores rather than boosting student performance. We also expect that students who were willing to take the CBT likely felt comfortable using computers. Such students were more likely to have come from relatively well-to-do families. By the inclusion of the variable “economically disadvantaged,” computer skills were indirectly taken into consideration. Therefore, omission of the variable related to computer skills in our propensity score estimation would be unlikely to alter the mode comparison results.

Limitations and Future Research

This study only considered the Rasch/1PL model in combination with the PCM for calibration, and a fixed parameter approach as the equating method. Other IRT models such as the two-parameter logistic model in combination of the Generalized Partial Credit models and other equating methods may be used for future studies. It is worth investigating if what has been found in this article with the three matching approaches would remain consistent when different IRT models and different equating methods are used.

Another area for further investigation is the impact of imputation on mode comparison results. As mentioned above, we only used one set of imputed values for each variable, and did not use the full capacity of the multiple imputation package (i.e., did not use the multiple set of imputed values). Future research could involve analyzing and including each set of imputed values to evaluate how this would affect comparability results.

Propensity score approaches (including propensity score matching) are still relatively new in the field of mode comparison studies. Because of the current trend of shifting all tests to CBT, it is expected that many comparability studies will be conducted within the next few years as both test delivery modes are expected to coexist for some time. Such test delivery modes’ coexistence is driven by schools needing time to meet various technology and infrastructure requirements. We consider comparability studies to be more meaningful if they are conducted before item calibrations and conversion tables are finalized. Moreover, as demonstrated in this study, different matching approaches lead to differing degrees of matching quality, which in turn may result in different conclusions with regard to mode comparability. Therefore, we recommend using the OPM-OBT approach as delineated in this article for forming comparable samples.

Acknowledgments

The authors would like to thank Joseph Martineau for comments he provided that helped improve this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. In particular, we considered the number of English language learners (ELLs) per school building, as the testing program under current mode comparison investigation is for ELLs only.
2. Math and reading were chosen here not only because they are the most consistently administered tests across the grade levels under consideration but also because performance on these two subject tests are considered in decision making for these ELLs.

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, 6(9). Retrieved from <http://www.jtla.org>
- Bowers, J., Fredrickson, M., & Hansen, B. (2010). RItools: Randomization inference tools. R package version 0.1-11 [Computer software]. Retrieved from <http://www.jakebowers.org/RItools.html>
- Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1), 31-72.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14, 464-504.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233-255.
- Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99, 609-618.
- Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23, 219-236.
- Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15, 609-627.
- Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, 15, 234-249.
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8), 1-28.

- Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, 101, 901-910.
- Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424-453.
- Keng, L., McClarty, K. L., & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas Assessment of Knowledge and Skills. *Applied Measurement in Education*, 21, 207-226.
- Kim, D.-H., & Huynh, H. (2007). Comparability of computer and paper-and-pencil versions of algebra and biology assessments. *Journal of Technology, Learning, and Assessment*, 6(4). Retrieved from <http://www.jtla.org>
- 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, 554-570.
- Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 population: A synthesis. *Applied Measurement in Education*, 22, 22-37.
- Linacre, J. M. (2009). WINSTEPS (Version 3.68.2) [Computer software]. Chicago, IL: Winsteps.com.
- Lottridge, S. M., Nicewander, W. A., & Mitzel, H. C. (2011). A comparison of paper and online tests using a within-subjects design and propensity score matching study. *Multivariate Behavioral Research*, 46, 544-566.
- Muthén, L. K., & Muthén, B. O. (2012). Mplus (Version 7.11) [Computer software]. Los Angeles, CA: Muthén & Muthén.
- Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement*, 22, 71-83.
- Puhan, P., Boughton, K., & Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerized testing. *Journal of Technology, Learning, and Assessment*, 6(3). Retrieved from <http://www.jtla.org>
- Randall, J., Sireci, S., Li, X., & Kaira, L. (2012). Evaluating the comparability of paper- and computer-based science tests across sex and SES subgroups. *Educational Measurement: Issues and Practice*, 31(4), 2-12.
- Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-524.
- Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33-38.
- Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement*, 71, 849-869.
- Sheskin, D. J. (2011). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL: Chapman & Hall/CRC.
- Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15, 250-267.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25, 1-21.

- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-67.
- Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006, April). *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.
- Way, W. D., Lin, C.-H., & Kong, J. (2008, March). *Maintaining score equivalence as test transition online: Issues, approaches and trends*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.