Appendix A—COMPARABILITY STUDY

MEPA

COMPARABILITY STUDY

Measured Progress

MEPA COMPARABILITY STUDY

I. Introduction

The Massachusetts English Proficiency Assessment (MEPA) program assesses Limited English Proficient (LEP) students in grades K through 12 to determine whether they have achieved sufficient proficiency to seriously consider removing their LEP status. To this end, MEPA assesses English proficiency in four domains: speaking, listening, reading, and writing. The speaking and listening components are assessed through observation in the classroom setting. The reading and writing components are assessed by fixed test forms that employ a combination of multiple-choice and open-response items. The study reported here focuses on only the reading and writing components of MEPA. For simplicity we will refer to these two components as "MEPA," but the reader should keep in mind that MEPA in actuality consists of all four components.

For the MEPA administration in the Spring of 2010, two versions of the MEPA tests were employed in grades 3 to 12. One version was a paper-based test (PBT) that was administered to the vast majority of students, and the other version was a computer-based test (CBT) that was administered at a limited number of schools that volunteered to have their students assessed in this way. For purposes of our analyses, we obtained complete records for 31,192 PBT students and 4247 CBT students – the combined total represents approximately 99.9% of all the MEPA test-takers in grades 3 to 12. Thus, approximately 12% of the MEPA student test-takers evaluated took the CBT version. The CBT version was introduced in the Spring 2010 administration as part of a gradual multi-year transition of the MEPA program from PBT to CBT. As part of this transition, the current study was carried out to investigate the comparability of the PBT and CBT versions.

II. Propensity score matching

When large samples are employed, as in the current study, the best method for conducting a comparability study would be to randomly assign students to the PBT and CBT groups. However, the only way that a sufficient sample size could be obtained for the CBT group was to allow every school to take the CBT that volunteered to take it.

When random assignment cannot be employed, as in the current study, an alternative method is a match-pairs design. In this design, each member of one group is matched with a member of the second group on a set of variables (called covariates) that are considered to be possible important influences on the variable of interest – in our case, performance on the MEPA. Sometimes finding exact matches on the covariates is difficult; and, in this case, propensity score matching (Rudner & Peyton, 2006; Rosenbaum & Rubin, 1985; Rubin, 1997; Joffe & Rosenbaum, 1999) can provide an effective alternative. In propensity score matching, discriminant function or logistic regression analysis is used to find the linear combination of the covariates that best discriminates between the two groups. This linear combination of the covariates is called a propensity score. Then members of the two groups are matched on propensity score, and a matched-pairs analysis is carried out. Details specific to the current study are given below.

III. Methods

Data

For each MEPA test form administered in grades 3 to 12, there are three assessment sessions for reading and another three sessions for writing, but students only take two sessions of each. In particular, students who have been identified as having lower levels of proficiency in reading are guided by their teachers to take Sessions 1 and 2 of reading, while students having relatively higher levels of proficiency in reading are guided to take Sessions 2 and 3 of reading. The exact same process is repeated for the writing test. Because reading and writing proficiency are highly correlated with each other, over 90% of the students take the same sessions in both reading and writing. In other words, the vast majority of students take either Sessions 1 and 2 in both reading and writing or take Sessions 2 and 3 in both reading and writing. For simplicity, the analyses in the current study focus on these students, the ones who took the same sessions in both reading and writing.

Another notable feature of the MEPA program is that multiple grades are clustered together into “grade spans” for test administration purposes, namely K-2, 3-4, 5-6, 7-8, and 9-12. These grade-spans are used because (a) the ELL skills assessed in the grades within a given grade-span are considered similar enough to be assessed by a single test and (b) the use of grade-spans permits direct measurement of progress across the grades within a grade-span. As previously noted, although MEPA is administered in grades K to 12, the CBT was only administered in grades 3 to 12, so that our analysis is thus restricted to the corresponding grade-spans. Although a separate test is administered for each grade-span, the raw scores for every test are scaled to a range of 400 to 550. Similarly, separate scaled scores for reading and writing, on a scale of 0 to 30, are also provided. Still, these scaled scores are not intended to be comparable between grade-spans since there are no common items or students across grade-spans for a given administration.

Analysis

Comparison groups. Instead of doing a separate analysis for each grade-span, we combined students across grade-spans to form the two groups to provide the most powerful analysis. Thus, the CBT group is defined as the union across all grade-spans of all students who took the CBT. Similarly, the PBT group is the union of all the students who took the PBT.

Variable of interest. Three variables of interest were defined for the current study, namely, the MEPA scaled score for reading and writing combined, the separate reading scaled score, and the separate writing scaled score. Although these scale scores are not comparable across grade-spans, they do provide the convenience of metrics that are immediately recognizable and interpretable to all parties of interest associated with the MEPA program.

Covariates. As recommended by the MEPA Technical Advisory Committee (TAC), two covariates were used for propensity score matching: (a) grade level and (b) score on the English Language Arts (ELA) test of the Massachusetts Comprehensive Assessment

System (MCAS). Approximately 82% of the MEPA CBT students (and approximately 76% of the MEPA PBT students) took MCAS in addition to MEPA. The matching score used for ELA was the raw score on the common multiple-choice items.

In addition to this primary analysis using the recommended covariates, further secondary analyses using additional covariates were also conducted to buttress the original recommended analysis with additional validity evidence. The additional covariates included gender, economic status, and native language.

An additional 3.5 percent of the students who took the CBT MEPA in Spring 2010 were newly enrolled LEP students who did not have MCAS ELA scores but did have MEPA scores from the Fall MEPA test that they took soon after enrolling. Thus, out of an abundance of caution, one other secondary analysis is also reported here in which these students were also included using all the covariates listed above but with their Fall MEPA scores being used in place of MCAS ELA scores.

Propensity score matching. A logistic regression analysis was conducted to find the linear combination of the covariates that best distinguished membership in the two groups. Because the PBT was the much larger group, the analysis proceeded by finding members of the PBT group that perfectly matched members of the CBT group in terms of propensity score. When multiple members of the PBT group provided a perfect match with a CBT group member, one of these PBT members was randomly selected for matching purposes.

Effect size calculation. After matching the two groups on propensity score, the mean and the standard deviation of each variable of interest (MEPA scaled score for reading and writing combined, MEPA scaled score for reading, and MEPA scaled score for writing), was calculated for the matched groups. Cohen's (1992) effect size was then calculated on the difference between the two groups for each variable of interest.

IV. Results

Primary Analysis

First, in Table 1 we provide descriptive statistics on the two groups, prior to doing any matching. In particular, we provide the effect size difference between the two groups using MEPA scaled score as well as the separate scaled scores for reading and writing. These effect sizes are provided merely as a baseline for comparison. At this point without having done further analysis yet, it is not known whether the two groups are matched well on the covariates. Table 1 shows an effect size of 0.25 to 0.33 in favor of the PBT group, meaning that the PBT group performed better on MEPA than did the CBT group, although the difference is considered small according to Cohen (1992). This difference may become either bigger or smaller, after a matching sample is extracted from the PBT group to compare with the CBT group.

Table 1 Comparison of scaled scores between CBT & PBT without propensity score matching

	CBT			PBT			
	N	Mean	S.D.	N	Mean	S.D.	Effect size
MEPA Overall Scaled Score	4247	476.44	24.96	31192	482.95	23.84	0.27
MEPA Reading Scaled Score	4247	13.88	5.07	31192	15.59	5.15	0.33
MEPA Writing Scaled Score	4247	14.50	5.20	31192	15.75	4.94	0.25

Table 2 provides a comparison of the PBT and CBT groups in terms of the two covariates, MCAS ELA score on the multiple-choice items and the distribution of the groups across the grade levels. The ELA scores within each grade level were standardized based on the mean and standard deviation of the scores for the two groups combined within each grade level. Then, the overall average of these standardized scores was used to describe each group and to calculate the effect size between them. Notice that the sample sizes are smaller in Table 2 in comparison to Table 1 because not all MEPA students took the MCAS. Table 2 clearly indicates that the PBT group has higher ELA scores with a positive effect size of 0.14. Table 2 also shows that there are notable differences on how the two groups are distributed across the grade levels. Because the difference in ELA score shown in Table 2 is in the same direction as the effect size in Table 1, matching on ELA score will obviously reduce the effect size difference between the two groups. It is not obvious how matching the distribution across grade levels will influence effect size.

Table 2 Comparison of covariates between groups

		CBT			PBT			
		N	Mean	S.D.	N	Mean	S.D.	Effect size
MCAS ELA MC Raw Score(z)		3480	-0.12	0.98	23765	0.02	1.00	0.14
		N	%		N	%		
Grade Level	3	686	20		5096	21		
	4	593	17		4732	20		
	5	464	13		3811	16		
	6	553	16		2994	13		
	7	471	14		2596	11		
	8	515	15		2425	10		
	10	198	6		2111	9		

Next, propensity score matching was conducted using MCAS ELA score and grade level as covariates. Members of the PBT group were selected in the manner described above so that they matched the propensity scores of each of the CBT group members. Table 3 describes how well the two groups are matched on the covariates. The results show that the matching is perfect.

Table 3 Comparison of covariates between groups after matching

		CBT			PBT			
		N	Mean	S.D.	N	Mean	S.D.	Effect size
MCAS ELA MC Raw Score(z)		3480	-0.12	0.98	3480	-0.12	0.98	0.00
		N	%		N	%		
Grade Level	3	686	20		686	20		
	4	593	17		593	17		
	5	464	13		464	13		
	6	553	16		553	16		
	7	471	14		471	14		
	8	515	15		515	15		
	10	198	6		198	6		

After matching on propensity score for these two covariates, the two groups are compared again in Table 4 on the variables of interest, MEPA total scaled score, reading scaled score, and writing scaled score. The results show that the effect sizes have now been reduced to a range of 0.19 to 0.25.

Table 4 Comparison of scaled scores between groups after matching

	CBT			PBT			
	N	Mean	S.D.	N	Mean	S.D.	Effect size
MEPA Overall Scaled Score	3480	476.72	25.62	3480	481.77	23.86	0.20
MEPA Reading Scaled Score	3480	13.99	5.23	3480	15.29	5.22	0.25
MEPA Writing Scaled Score	3480	14.56	5.31	3480	15.53	5.03	0.19

Secondary Analysis: Additional Covariates

As described above, a secondary analysis was conducted requiring students to be matched on more covariates in addition to those used in the primary analysis. The additional covariates are gender, economically disadvantaged (labeled as “EconDis” in the table; dichotomously coded as 1 if the characteristic pertained to the student, 0 otherwise), and primary language (dummy coded for six languages). Table 5 provides a comparison of the PBT and CBT groups in terms of all the covariates prior to doing any matching. The two groups are seen to have the same gender percentages, a small but notable difference in percent economically disadvantaged and small differences in the distribution across the six languages. Note that the sample size in Table 5 is the same as Table 2 which means that the six languages comprise all the languages that were present for the students in Table 2.

Next, propensity score matching was conducted using the all the covariates. As in the primary analysis, members of the PBT group were selected so that they matched the propensity scores of each of the CBT group members. There were some members of the CBT group who had propensity scores that could not be matched with anyone in the PBT group. This resulted in the sample size being reduced from 3480 in the primary analysis to 2880 in this secondary analysis. Table 6 describes how well the two groups were matched on this expanded set of covariates. The results show that the matching was again perfect.

After matching on propensity score for this expanded set of covariates, the two groups were compared again in Table 7 on the variables of interest – MEPA total scaled score, reading scaled score, and writing scaled score. The results show that the effect sizes changed only slightly and still range from 0.19 to 0.25.

Table 5 Comparison of expanded covariates between groups

		CBT			PBT			
		N	Mean	S.D.	N	Mean	S.D.	Effect size
MCAS ELA MC Raw Rcore(z)		3480	-0.12	0.98	23765	0.02	1.00	0.14
		N	%		N	%		
Grade Level	3	686	20		5096	21		
	4	593	17		4732	20		
	5	464	13		3811	16		
	6	553	16		2994	13		
	7	471	14		2596	11		
	8	515	15		2425	10		
	10	198	6		2111	9		
Gender	Female	1993	47		14502	47		
	Male	2223	53		16515	53		
EconDis	Yes	3792	90		25683	83		
Language	Spanish	2872	77		15567	63		
	Portuguese	208	6		2005	8		
	Cape Verdean	251	7		1574	6		
	Haitian Creole	162	4		1822	7		
	Khmer/Khmai	108	3		1327	5		
	Vietnamese	74	2		1200	5		
	Chinese	54	1		1148	5		

Table 6 Comparison of expanded covariates between groups after matching

		CBT			PBT			
		N	Mean	S.D.	N	Mean	S.D.	Effect size
MCAS ELA MC Raw Score(z)		2880	-0.16	0.96	2880	-0.16	0.96	0.00
		N	%		N	%		
Grade Level	3	586	20		586	20		
	4	516	18		516	18		
	5	403	14		403	14		
	6	478	17		478	17		
	7	345	12		345	12		
	8	394	14		394	14		
	10	158	5		158	5		
Gender	Female	1386	48		1386	48		
	Male	1494	52		1494	52		
EconDis	Yes	2704	94		2704	94		
Language	Spanish	2305	80		2305	80		
	Portuguese	111	4		111	4		
	Cape Verdean	182	6		182	6		
	Haitian Creole	114	4		114	4		
	Khmer/Khmai	86	3		86	3		
	Vietnamese	49	2		49	2		
	Chinese	33	1		33	1		

Table 7 Comparison of two groups after matching on expanded covariates

	CBT			PBT			
	N	Mean	S.D.	N	Mean	S.D.	Effect size
MEPA Overall Scaled Score	2880	475.80	24.77	2880	481.31	23.71	0.23
MEPA Reading Scaled Score	2880	13.80	5.05	2880	15.08	5.02	0.25
MEPA Writing Scaled Score	2880	14.37	5.18	2880	15.32	5.05	0.19

Secondary Analysis: Additional Covariates and Extended Score Matching

As described above, another analysis was conducted using the expanded list of covariates but allowing the score matching to include the score from the Fall MEPA test when the MCAS ELA score was not available for a student. This resulted in an additional 147 CBT students and an additional 1374 PBT students to be included in the secondary analysis. Table 8 provides descriptive statistics on the covariates for the two groups. The scores on the Fall MEPA test have been standardized by the mean and standard deviation of the scaled test scores within each grade level.

Next, propensity score matching was conducted using the all the covariates, including the extended score covariate. Again, members of the PBT group were selected so that they matched the propensity scores of each of the CBT group members. The matching resulted in the loss of only one member of the CBT group who had a propensity score that could not be matched with anyone in the PBT group. This resulted in a sample size of 3026, an increase of 146 over the above secondary analysis. Table 9 describes how well the matching was accomplished for the two groups and, once again, the matching was perfect.

After matching on propensity score for the expanded set of covariates with the extended score covariate, the two groups were compared again in Table 10 on the variables of interest – MEPA total scaled score, reading scaled score, and writing scaled score. The results show that the effect sizes again changed only slightly and still range from 0.16 to 0.24.

Table 8. Comparison of covariates between groups with extended score matching

		CBT			PBT			
		N	Mean	S.D.	N	Mean	S.D.	Effect size
MCAS ELA MC Raw Score(z)		3480	-0.12	0.98	23765	0.02	1.00	0.14
MEPA(Fall) Scaled Score(z)		147	-0.12	1.03	1374	0.01	1.00	0.13
		N	%		N	%		
Grade Level	3	686	20		5096	21		
	4	593	17		4732	20		
	5	464	13		3811	16		
	6	553	16		2994	13		
	7	471	14		2596	11		
	8	515	15		2425	10		
	10	198	6		2111	9		
Gender	Female	1993	47		14502	47		
	Male	2223	53		16515	53		
EconDis	Yes	3792	90		25683	83		
Language	Spanish	2872	77		15567	63		
	Portuguese	208	6		2005	8		
	Cape Verdean	251	7		1574	6		
	Haitian Creole	162	4		1822	7		
	Khmer/Khmai	108	3		1327	5		
	Vietnamese	74	2		1200	5		
	Chinese	54	1		1148	5		

Table 9. Comparison of expanded covariates with extended score after matching

		CBT			PBT			
		N	Mean	S.D.	N	Mean	S.D.	Effect size
MCAS ELA MC Raw Score(z)		2880	-0.16	0.96	2880	-0.16	0.96	0.00
MEPA(Fall) Scaled Score(z)		146	-0.11	1.03	146	-0.11	1.03	0.00
		N	%		N	%		
Grade Level	3	586	20		586	20		
	4	516	18		516	18		
	5	403	14		403	14		
	6	478	17		478	17		
	7	345	12		345	12		
	8	394	14		394	14		
	10	158	5		158	5		
Gender	Female	1386	48		1386	48		
	Male	1494	52		1494	52		
EconDis	Yes	2704	94		2704	94		
Language	Spanish	2305	80		2305	80		
	Portuguese	111	4		111	4		
	Cape Verdean	182	6		182	6		
	Haitian Creole	114	4		114	4		
	Khmer/Khmai	86	3		86	3		
	Vietnamese	49	2		49	2		
	Chinese	33	1		33	1		

Table 10. Comparison of groups using expanded covariates with extended score

	CBT			PBT			
	N	Mean	S.D.	N	Mean	S.D.	Effect size
MEPA (Spring) Scaled Score	3026	475.41	24.77	3026	480.18	23.66	0.20
MEPA (Spring) Reading Scaled Score	3026	13.74	5.02	3026	14.94	4.96	0.24
MEPA (Spring) Writing Scaled Score	3026	14.32	5.18	3026	15.12	5.03	0.16

V. Concluding Remarks

The MEPA program has embarked on a multi-year transition from a PBT to a CBT and has completed the first year of that transition. As part of this transition a study has been conducted to evaluate the comparability of the student test-taking experience between these two environments. Because the CBT group, unlike the PBT group, consisted of self-selected volunteers, the comparability study was conducted using a subsample of the PBT group that was matched with the CBT group on relevant covariates. Using these matched groups, an effect size difference was calculated between the two groups. Three effect sizes, based on three standard reported MEPA scores, were calculated; and these effect sizes ranged from 0.19 to 0.25.

As a validity check, two follow-up analyses were conducted using an expanded list of covariates. Both of these analyses gave effect sizes that were nearly identical to the effect sizes from the original analysis

These effect sizes are small and do not warrant treating the CBT and PBT scores as though they come from different tests. In particular, we conclude that no equating of the CBT and PBT scores is necessary for the Spring 2010 MEPA administration.

VI. References

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-129.

Joffe, M. M., & Rosenbaum, P. R. (1999). Propensity scores. *American Journal of Epidemiology, 150,* 327-333.

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.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. [Supplement]. *Annals of Internal Medicine, 127* (8S), 757-763.

Rudner, L. M., & Peyton, J. (2006). Consider propensity scores to compare treatments. *GMAC Research Reports, RR-06-07.* GMAC: McLean, Virginia.