APPENDIX A—2012 & 2011 MEPA COMPARABILITY STUDIES

MEPA

COMPARABILITY STUDY

June 2012

Louis A. Roussos

Won Suk Kim

Jennifer Dunn

100 EDUCATION WAY, DOVER, NH 03820 (800) 431-8901
WWW.MEASUREDPROGRESS.ORG

1. INTRODUCTION

The Massachusetts English Proficiency Assessment (MEPA) program assesses students in grades K–12 who are designated as English language learners (ELL)[1], and is used as one factor in determining whether they are ready to transition out of ELL status. MEPA assesses English proficiency in four domains: speaking, listening, reading, and writing. The speaking and listening components are assessed through a locally administered observational assessment. Reading and writing are assessed using fixed test forms that employ a combination of multiple-choice and constructed-response items. This study focuses on the comparability of paper-based and computer-based administrations for the reading and writing components of MEPA.

Prior to the spring 2010 administration, MEPA was administered solely as a paper-based test (PBT). For the spring 2010 MEPA administration, a computer-based test (CBT) was introduced in grades 3–12 in a limited number of schools on a voluntary basis. The CBT version was introduced as part of a gradual multi-year transition of the MEPA program from PBT to CBT. As part of this transition, a comparability study was conducted after the spring 2010 administration to investigate the comparability of the PBT and CBT versions. (See Appendix A: Comparability Study in the *Massachusetts English Proficiency Assessment 2011 Technical* Report for details of the 2010 comparability study.) The results of that study indicated that the PBT and CBT versions were sufficiently comparable that equating the two versions was not necessary. The study was repeated after the spring 2011 administration with the results again supporting the comparability of the PBT and CBT.

The transition from PBT to CBT continued in spring 2012. Both the PBT and CBT versions of MEPA were administered. The purpose of this study is to continue to monitor the comparability of the PBT and CBT versions of MEPA in the final year of the MEPA administration. In preparation for our analyses, we obtained complete records for 36,365 PBT students and 5,877 CBT students. Thus, approximately 13.9% of the MEPA student test-takers took the CBT version.

[1] In prior reports the term "Limited English proficient (LEP)" was used to describe students whose first language is a language other than English and who cannot perform ordinary class work in English. This term has now been replaced with "English language learners (ELL)."

2. PROPENSITY SCORE MATCHING

Ideally, for a comparability study, students are randomly assigned to one of the two study conditions, ensuring that the two groups are randomly equivalent. Instead, because the CBT group consisted of students from volunteer schools, a matched-pairs design was used to identify an equivalent group of students who took the PBT version. 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 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 then conducted. Details specific to the current study are given below.

3. METHODS

3.1. Data

For each MEPA test form administered in grades 3–12, there are three assessment sessions for reading and another three sessions for writing. Students are assigned to take only two sessions of each, based on their level of English proficiency. Students identified as having lower levels of proficiency in reading take Sessions 1 and 2 of the reading test (denoted as "r12" in the tables that follow), while students with intermediate or high levels of proficiency in reading take Sessions 2 and 3 (denoted as "r23" in the tables). The same process is repeated for the writing test (with sessions denoted by "w12" and "w23" in the tables). Thus, there are four different combinations of reading and writing sessions, each of which is regarded as a separate test form on which a separate raw-score to scale-score conversion table was required. We refer to each of these combinations as a "scale form" in the tables. However, 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 either take Sessions 1 and 2 in both reading and writing or take Sessions 2 and 3 in both reading and writing.

Another 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. Since the

CBT was only administered in grades 3–12, our analysis is restricted to the corresponding grade spans and excludes K–2. Students receive a scaled reading score from 0–30, a scaled writing score from 0–30, and a combined scaled score from 400–550. However, scaled scores are not comparable between grade spans since there are no common items or students tested across grade spans for any administrations.

3.2. Analysis

Comparison Groups. Instead of doing a separate analysis for each grade span, we combined students across grade spans into two groups: those who took the CBT and those who took the PBT.

Variables of Interest. Three variables of interest were defined for the current study: the MEPA combined scaled score for reading and writing, the separate reading scaled score, and the separate writing scaled score. Although these scaled scores are not comparable across grade spans, they do provide the convenience of metrics that are recognizable and interpretable.

Covariates. Based on our experience with the comparability studies for the spring 2010 and the spring 2011 MEPA, as well as discussions with the Massachusetts Department of Elementary and Secondary Education (ESE) and the MCAS Technical Advisory Committee (TAC), three variables were chosen as the primary covariates for propensity score matching: (1) grade level, (2) score on the English Language Arts (ELA) test of the Massachusetts Comprehensive Assessment System (MCAS), and (3) the combination of reading and writing sessions to which a student was assigned. Approximately 47% of the MEPA CBT students and approximately 51% of the MEPA PBT students had official MCAS ELA scores[2]. The matching score used was the MCAS ELA scaled score.

A secondary analysis using additional covariates was then conducted to provide additional validity evidence. The additional covariates were gender, economic status, and native language.

An additional 14% of the students who took the CBT MEPA in spring 2012 were newly enrolled English language learner (ELL) students who did not have MCAS ELA scores, but did have MEPA scores from the previous fall MEPA test administration. Another analysis was conducted in which these students were included using the primary and secondary covariates listed above, but with their fall MEPA scores 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

[2] These percentages reflect two factors: (1) students reported to the Department as first-year ELL students are not required to take the MCAS ELA test, and (2) at the high school level, MCAS ELA is administered at grade 10 only.

PBT was the 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 by propensity score, the mean and the standard deviation of each variable of interest (MEPA combined scaled score for reading and writing, 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.

4. RESULTS

4.1. Primary Analysis

In Table 1, we provide descriptive statistics on the two groups prior to matching. In particular, we provide the effect size difference between the two groups using MEPA combined scaled scores as well as the separate scaled scores for reading and writing. These effect sizes are provided as a baseline for comparison. We do not know whether the two groups are matched well on the covariates without further analysis. Table 1 shows effect sizes of 0.29 to 0.32 of a standard deviation 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 change once a matching sample is extracted from the PBT group to compare with the CBT group.

Table 1. Comparison of Scaled Score between CBT & PBT without Propensity Score Matching

		CBT			*PBT*			*Effect Size*
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MEPA	Combined Scaled Score	5,877	480.8	24.6	36,365	487.7	23.5	0.29
	REA Scaled Score	5,877	14.8	5.2	36,365	16.4	5.0	0.31
	WRT Scaled Score	5,877	15.3	5.0	36,365	16.9	5.0	0.32

Table 2 provides a comparison of the PBT and CBT groups in terms of the three covariates: MCAS ELA scaled score, grade level, and the reading and writing sessions to which students were assigned (denoted as "scale form" in Table 2). 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. The average of these standardized scores was used to describe each group

and to calculate the effect size between them. The sample sizes are smaller in Table 2 than in Table 1 because (as mentioned in footnote 1 on page 3) students reported to the Department as first-year ELL students are not required to take the MCAS ELA and because, at the high school level, MCAS ELA is only administered at grade 10. Table 2 clearly indicates that the PBT group has higher ELA scores with a positive effect size of 0.16. Table 2 also shows that the differences in how the two groups are distributed across the grade levels are small, but the differences in their distribution across the reading and writing sessions are significant. Because the difference in ELA scores shown in Table 2 is in the same direction as the effect size in Table 1, matching by ELA scores will reduce the effect size between the two groups. Because students who are assigned to sessions r23/w23 likely have higher levels of achievement than those assigned to r12/w12, the differences in "scale form" distribution again favor PBT, thus implying that the "scale form" matching will also reduce the effect size in Table 1.

Table 2. Comparison of Covariates between Groups

		CBT			*PBT*			*Effect Size*
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MCAS 2012	ELA Scaled Score (z)	2,778	-0.14	0.95	18,515	0.02	1.01	0.16
		CBT			*PBT*			
		N	*%*		*N*	*%*		
Grade Level	3	649	19		4,963	24		
	4	629	19		3,915	19		
	5	525	16		3,416	16		
	6	487	15		2,615	13		
	7	438	13		2,307	11		
	8	400	12		1,859	9		
	10	209	6		1,728	8		
Scale Form	r12/w23	61	1		297	1		
	r12/w12	2,027	35		9,830	27		
	r23/w23	3,686	63		25,936	71		
	r23/w12	103	2		302	1		

Next, propensity score matching was conducted using MCAS ELA score, grade level, and scale form as covariates. Members of the PBT group were selected in the manner described above to match the propensity scores of each of the CBT group members. Table 3 demonstrates how well the two groups are matched on the covariates. The effect size of 0.0 indicates perfect matching.

Table 3. Comparison of Covariates between Groups after Matching

		CBT			PBT			Effect Size
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MCAS 2012	ELA Scaled Score (z)	2,757	-0.14	0.95	2,757	-0.14	0.95	0.00
		CBT			*PBT*			
		N	*%*		*N*	*%*		
Grade Level	3	565	21		565	21		
	4	541	20		541	20		
	5	440	16		440	16		
	6	401	15		401	15		
	7	336	12		336	12		
	8	302	11		302	11		
	10	172	6		172	6		
Scale Form	r12/w23	13	1		13	1		
	r12/w12	460	17		460	17		
	r23/w23	2,259	82		2,259	82		
	r23/w12	22	1		22	1		

After matching on propensity score for these three covariates, the two groups were compared on the variables of interest: MEPA combined scaled score, reading scaled score, and writing scaled score. The results in Table 4 show that the effect sizes have now been reduced to a range of 0.13–0.20. The effect size of the MEPA combined scaled score (the one we're most concerned with) is 0.13, indicating that the CBT and PBT are comparable assessments.

Table 4. Comparison of Scaled Scores between Groups after Matching

		CBT			PBT			Effect Size
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MEPA	Combined Scaled Score	2,757	491.8	18.2	2,757	494.1	18.2	0.13
	REA Scaled Score	2,757	16.7	4.5	2,757	17.3	4.5	0.13
	WRI Scaled Score	2,757	17.3	3.9	2,757	18.0	4.0	0.20

4.2. Secondary Analysis: Additional Covariates

As described above, a secondary analysis was conducted requiring students to be matched on the additional covariates of gender, economically disadvantaged status (labeled as "EconDis" in the table; dichotomously coded as 1 if the characteristic pertained to the student, 0 otherwise), and primary language. Table 5 provides a comparison of the PBT and CBT groups in terms of all the covariates (primary and secondary) prior to doing matching. The two groups are seen to have small differences in gender percentages and in percent economically disadvantaged. More noticeable differences can be seen in the distribution across the seven languages. While both groups have

Spanish as the predominant language, the PBT group has a substantially higher percentage (81%) than the CBT group (63%). This difference of 18 percentage points is primarily accounted for by the CBT group having 11% more students whose native languages were reported as Cape Verdean (a Portuguese-based language) and Haitian Creole (a French-based language). We had no prior hypothesis about how adjusting for these language differences might affect the results. Note that the sample size in Table 5 is the same as that in Table 2, indicating that the seven languages comprise all the languages used by 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 to match the propensity scores of each CBT group member. Some members of the CBT group had propensity scores that could not be matched with any members of the PBT group. This resulted in the sample size being reduced from 5,514 in the primary analysis to 4,302 in this secondary analysis. Table 6 describes how well the two groups were matched on this expanded set of covariates. The results again show perfect matching.

After matching on propensity score for this expanded set of covariates, the two groups were compared in Table 7 on the variables of interest: MEPA combined scaled score, reading scaled score, and writing scaled score. The results show that the effect sizes changed only slightly, ranging from 0.13 to 0.25.

Table 5. Comparison of Expanded Covariates between Groups

		CBT			*PBT*			*Effect Size*
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MCAS 2012	ELA Scaled Score (z)	2,778	-0.14	0.95	18515	0.02	1.01	0.16
		CBT			*PBT*			
		N	*%*		*N*	*%*		
Grade Level	3	649	19		4,963	24		
	4	629	19		3,915	19		
	5	525	16		3,416	16		
	6	487	15		2,615	13		
	7	438	13		2,307	11		
	8	400	12		1,859	9		
	10	209	6		1,728	8		
Scale Form	r12/w23	61	1		297	1		
	r12/w12	2,027	35		9,830	27		
	r23/w23	3,686	63		25,936	71		
	r23/w12	103	2		302	1		
Gender	Female	2,807	48		16,811	46		
	Male	3,070	52		19,482	54		
EconDis	Yes	4,923	84		29,175	81		
Language	Spanish	3,994	81		18,153	63		
	Portuguese	294	6		1,832	6		
	Cape Verdean	109	2		2,294	8		
	Haitian Creole	195	4		2,659	9		
	Khmer/Khmai	139	3		1,388	5		
	Vietnamese	152	3		1,345	5		
	Chinese	73	1		1,179	4		

Table 6. Comparison of Expanded Covariates between Groups after Matching

		CBT			PBT			Effect Size
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MCAS 2012	ELA Scaled Score (z)	2,151	-0.25	0.84	2,151	-0.25	0.84	0.00
		CBT			*PBT*			
		N		*%*	*N*		*%*	
Grade Level	3	471		22	471		22	
	4	418		19	418		19	
	5	337		16	337		16	
	6	323		15	323		15	
	7	248		12	248		12	
	8	225		10	225		10	
	10	129		6	129		6	
Scale Form	r12/w23	5		0	5		0	
	r12/w12	338		16	338		16	
	r23/w23	1,798		84	1,798		84	
	r23/w12	10		0	10		0	
Gender	Female	1,037		48	1,037		48	
	Male	1,114		52	1,114		52	
EconDis	Yes	2,006		93	2,006		93	
Language	Spanish	1,853		86	1,853		86	
	Portuguese	68		3	68		3	
	Cape Verdean	35		2	35		2	
	Haitian Creole	71		3	71		3	
	Khmer/Khmai	60		3	60		3	
	Vietnamese	51		2	51		2	
	Chinese	13		1	13		1	

Table 7. Comparison of Scaled Scores Between Groups after Matching

		CBT			PBT			Effect Size
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MEPA	Combined Scaled Score	2,151	490.05	16.99	2,151	493.11	17.15	0.18
	REA Scaled Score	2,151	16.27	4.18	2,151	16.81	4.31	0.13
	WRI Scaled Score	2,151	16.97	3.78	2,151	17.91	3.85	0.25

4.3. 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 in cases where the MCAS ELA score was not available for a student. This process yielded an additional 819 CBT students and an additional 4,318 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.

Table 8. Comparison of Covariates between Groups with Extended Score Matching

		CBT			*PBT*			*Effect Size*
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MCAS 2012	ELA Scaled Score (z)	2,778	-0.14	0.95	18,515	0.02	1.01	0.16
MEPA Fall	Scaled Score (z)	819	-0.18	0.99	4,318	0.04	1.00	0.22
		CBT			*PBT*			
		N	*%*		*N*	*%*		
Grade Level	3	649	19		4,963	24		
	4	629	19		3,915	19		
	5	525	16		3,416	16		
	6	487	15		2,615	13		
	7	438	13		2,307	11		
	8	400	12		1,859	9		
	10	209	6		1,728	8		
Scale Form	r12/w23	61	1		297	1		
	r12/w12	2,027	35		9,830	27		
	r23/w23	3,686	63		25,936	71		
	r23/w12	103	2		302	1		
Gender	Female	2,807	48		16,811	46		
	Male	3,070	52		19,482	54		
EconDis	Yes	4,923	84		29,175	81		
Language	Spanish	3,994	81		18,153	63		
	Portuguese	294	6		1,832	6		
	Cape Verdean	109	2		2,294	8		
	Haitian Creole	195	4		2,659	9		
	Khmer/Khmai	139	3		1,388	5		
	Vietnamese	152	3		1,345	5		
	Chinese	73	1		1,179	4		

Next, propensity score matching was conducted using the all the covariates, including the extended score covariate. 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 30 members of the CBT group who had a propensity score that could not be matched with any members of the PBT group. This resulted in a sample size of 2,946, an increase of 795 over the original secondary analysis. Table 9 shows the matching of the two groups and again indicates perfect matching.

Table 9. Comparison of Expanded Covariates with Extended Score after Matching

		CBT			*PBT*			*Effect Size*
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MCAS 2012	ELA Scaled Score (z)	2,151	-0.25	0.84	2,151	-0.25	0.84	0.00
MEPA Fall	Scaled Score (z)	795	-0.18	0.97	795	-0.18	0.97	0.00
		CBT			*PBT*			
		N	*%*		*N*	*%*		
Grade Level	3	471	22		471	22		
	4	418	19		418	19		
	5	337	16		337	16		
	6	323	15		323	15		
	7	248	12		248	12		
	8	225	10		225	10		
	10	129	6		129	6		
Scale Form	r12/w23	1,037	48		1,037	48		
	r12/w12	1,114	52		1,114	52		
	r23/w23	2,006	93		2,006	93		
	r23/w12	1,853	86		1,853	86		
Gender	female	68	3		68	3		
	male	35	2		35	2		
EconDis	yes	71	3		71	3		
Language	Spanish	60	3		60	3		
	Portuguese	51	2		51	2		
	Cape Verdean	13	1		13	1		
	Haitian Creole	471	22		471	22		
	Khmer/Khmai	418	19		418	19		
	Vietnamese	337	16		337	16		
	Chinese	323	15		323	15		

After matching on propensity score for the expanded set of covariates with the extended score covariate, the two groups were compared on the variables of interest: MEPA combined scaled score, reading scaled score, and writing scaled score. The results shown in Table 10 indicate that the effect sizes were slightly reduced and now ranged from 0.13 to 0.18.

Table 10. Comparison of Groups Using Expanded Covariates with Extended Score

		CBT			*PBT*			*Effect Size*
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MEPA Spring	Scaled Score	2,946	482.88	23.61	2,946	485.85	23.90	0.13
	REA Scaled Score	2,946	15.11	5.00	2,946	15.77	5.08	0.13
	WRI Scaled Score	2,946	15.67	4.87	2,946	16.56	5.03	0.18

5. CONCLUDING REMARKS

The MEPA program has now completed three years of testing both computer-based and paper-based tests. As part of this process a study has been conducted each year to evaluate the comparability of the results from these two tests. 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 for the two groups. Three effect sizes, based on three standard reported MEPA scores, were calculated as follows: 0.13 for the MEPA combined scaled scores, 0.13 for the reading scaled scores, and 0.20 for the writing scaled scores.

To verify the validity of these results, two follow-up analyses were conducted using an expanded list of covariates. Both of these analyses gave effect sizes that were similar to the effect sizes from the original analysis. The most extensive analysis reported an effect size of 0.13 for MEPA combined scaled scores and 0.13 and 0.18 respectively for the reading and writing scaled scores.

These effect sizes fall below the threshold of 0.2 advised by the TAC, and thus do not warrant treating the CBT and PBT scores differently. Therefore, we recommend that no equating of the CBT and PBT scores occur for the spring 2012 MEPA administration.

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.

MEPA

COMPARABILITY STUDY

May 2011

Louis A. Roussos

Wonsuk Kim

Jennifer Dunn

100 EDUCATION WAY, DOVER, NH 03820 (800) 431-8901
WWW.MEASUREDPROGRESS.ORG

1. INTRODUCTION

The Massachusetts English Proficiency Assessment (MEPA) program assesses students in grades K–12 who are designated as limited English proficient (LEP) on their level of English proficiency, and is used as one factor in determining whether they are ready to transition out of LEP status. MEPA assesses English proficiency in four domains: speaking, listening, reading, and writing. The speaking and listening components are assessed through a locally administered observational assessment. Reading and writing are assessed using fixed test forms that employ a combination of multiple-choice and constructed-response items. This study focuses on the comparability of paper-based and computer-based administrations for the reading and writing components of MEPA.

Prior to the spring 2010 administration, MEPA was administered solely as a paper-based test (PBT). For the spring 2010 MEPA administration, a computer-based test (CBT) was introduced in grades 3–12 in a limited number of schools on a voluntary basis. The CBT version was introduced as part of a gradual multi-year transition of the MEPA program from PBT to CBT. As part of this transition, a comparability study was conducted after the spring 2010 administration to investigate the comparability of the PBT and CBT versions. (See Appendix A: Comparability Study in the *Massachusetts English Proficiency Assessment 2011 Technical Report* for details of the 2010 comparability study.) The results of that study indicated that the PBT and CBT versions were sufficiently comparable that equating the two versions was not necessary.

The transition from PBT to CBT continued in spring 2011. Both the PBT and CBT versions of MEPA were administered, with an increased number of students taking the CBT version, as planned. The purpose of this study is to continue to monitor the comparability of the PBT and CBT versions of MEPA in the second year. In preparation for our analyses, we obtained complete records for 34,834 PBT students and 6,254 CBT students. Thus, approximately 15.2% of the MEPA student test-takers took the CBT version.

2. PROPENSITY SCORE MATCHING

Ideally, for a comparability study, students are randomly assigned to one of the two study conditions, ensuring that the two groups are randomly equivalent. Instead, because the CBT group consisted of students from volunteer schools, a matched-pairs design was used to identify an equivalent group of students who took the PBT version. 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 MEPA. Sometimes finding exact matches on the covariates is difficult; 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. Members of the two groups are matched on propensity score, and a matched-pairs analysis is conducted. Details specific to the current study are given below.

3. METHODS

3.1 Data

For each MEPA test form administered in grades 3–12, there are three assessment sessions for reading and three assessment sessions for writing. Students are assigned to take only two sessions of each, based on their level of English proficiency. Students identified as having lower levels of proficiency in reading take Sessions 1 and 2 of the reading test (denoted as "r12" in the tables that follow), while students with intermediate or high levels of proficiency in reading take Sessions 2 and 3 (denoted as "r23" in the tables). The same process is repeated for the writing test (with sessions denoted by "w12" and "w23" in the tables). Thus, there are four different combinations of reading and writing sessions, each of which was regarded as a separate test form on which a separate raw-score to scale-score conversion table was required. We refer to each of these combinations as a "scale form" in the tables. However, because reading and writing proficiency are highly correlated with each other, over 90% of the students took the same set of sessions in both reading and writing. In other words, the vast majority of students either take Sessions 1 and 2 in both reading and writing or take Sessions 2 and 3 in both reading and writing.

Another 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. Since the CBT was only administered in grades 3–12, our analysis is restricted to the corresponding grade-spans and excludes K–2. Students receive a scaled reading score from 0–30, a scaled writing score from 0–30, and a combined scaled score from 400–550. However, scaled scores are not comparable between grade-spans since there are no common items or students tested across grade-spans for any administrations.

3.2 Analysis

Comparison Groups. Instead of doing a separate analysis for each grade-span, we combined students across grade-spans into two groups: those who took the CBT and those who took the PBT.

Variables of Interest. Three variables of interest were defined for the current study: the MEPA combined scaled score for reading and writing, the separate reading scaled score, and the separate writing scaled score. Although these scaled scores are not comparable across grade-spans, they do provide the convenience of metrics that are recognizable and interpretable.

Covariates. Based on our experience with the comparability study for the spring 2010 MEPA, as well as discussions with the Massachusetts Department of Elementary and Secondary Education (ESE) and the MCAS Technical Advisory Committee (TAC), three variables were chosen as the primary covariates for propensity score matching: (1) grade level, (2) scaled score on the English Language Arts (ELA) test of the Massachusetts Comprehensive Assessment System (MCAS), and (3) the combination of reading and writing sessions to which a student was assigned. Approximately 44% of the MEPA CBT students and 50% of the MEPA PBT students had official MCAS ELA scores[1]. The matching score used was the MCAS ELA scaled score.

A secondary analysis using additional covariates was then conducted to provide additional validity evidence. The additional covariates were gender, economic status, and native language.

An additional 18% of the students who took the CBT MEPA in spring 2011 were newly enrolled ELL students who did not have MCAS ELA scores, but did have MEPA scores from the previous fall MEPA test administration. Another analysis was conducted in which these students were included using the primary and secondary covariates listed above, but with their fall MEPA scores 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 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 by propensity score, the mean and the standard deviation of each variable of interest (MEPA combined scaled score for reading and

[1] These percentages reflect two factors: (1) students reported to the Department as first-year LEP students are not required to take the MCAS ELA test, and (2) at the high school level, MCAS ELA is administered at grade 10 only.

writing, 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.

4. RESULTS

4.1 Primary Analysis

In Table 1, we provide descriptive statistics on the two groups prior to matching. In particular, we provide the effect size difference between the two groups using MEPA combined scaled scores as well as the separate scaled scores for reading and writing. These effect sizes are provided as a baseline for comparison. We do not know whether the two groups are matched well on the covariates without further analysis. Table 1 shows effect sizes of 0.28 and 0.29 of a standard deviation in favor of the PBT group, meaning that the PBT group performed better on MEPA than the CBT group did, although the difference is considered small according to Cohen (1992). The effect size may change once a matching sample is extracted from the PBT group to compare with the CBT group.

Table 1. Comparison of Scaled Score between CBT & PBT without Propensity Score Matching

		CBT			*PBT*			*Effect Size*
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MEPA	Combined Scaled Score	6,254	478.3	25.6	3,4834	485.1	24.3	0.28
	REA Scaled Score	6,254	14.5	5.3	3,4834	16.0	5.4	0.29
	WRT Scaled Score	6,254	14.6	5.3	3,4834	16.1	5.0	0.29

Table 2 provides a comparison of the PBT and CBT groups in terms of the three covariates—MCAS ELA scaled score, grade level, and the reading and writing sessions students were assigned to (denoted as "scale form" in Table 2). 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. The average of these standardized scores was used to describe each group and to calculate the effect size between them. The sample sizes are smaller in Table 2 than in Table 1 because (as mentioned in footnote 1 on page 3) students reported to the Department as first-year LEP students are not required to take the MCAS ELA and because, at the high school level, MCAS ELA is only administered at grade 10. Table 2 clearly indicates that the PBT group has higher ELA scores with a positive effect size of 0.18. Table 2 also shows that the differences in how the two groups are distributed across the grade levels are small, but the differences in their distribution across the

reading and writing sessions are significant. Because the difference in ELA scores shown in Table 2 is in the same direction as the effect size in Table 1, matching by ELA scores will reduce the effect size between the two groups. Because students who are assigned to sessions r23/w23 likely have higher levels of achievement than those assigned to r12/w12, the differences in "scale form" distribution again favor PBT, thus implying that the "scale form" matching will also reduce the effect size in Table 1.

Table 2. Comparison of Covariates between Groups

		CBT			*PBT*			*Effect Size*
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MCAS 2010	ELA Scaled Score (z)	2,781	-0.15	0.95	17,432	0.02	1.01	0.18
		CBT			*PBT*			
		N		*%*	*N*		*%*	
Grade Level	3	693		21	4,552		23	
	4	689		21	4,053		20	
	5	454		14	3,025		15	
	6	489		15	2,436		12	
	7	434		13	2,166		11	
	8	310		9	1,878		9	
	10	234		7	1,750		9	
Scale Form	r12/w23	97		2	402		1	
	r12/w12	2,297		37	9789		28	
	r23/w23	3,809		61	24,327		70	
	r23/w12	51		1	316		1	

Next, propensity score matching was conducted using MCAS ELA score, grade level, and scale form as covariates. Members of the PBT group were selected in the manner described above to match the propensity scores of each of the CBT group members. Table 3 demonstrates how well the two groups are matched on the covariates. The effect size of 0.0 indicates perfect matching.

Table 3. Comparison of Covariates between Groups after Matching

		CBT			PBT			Effect Size
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MCAS 2010	ELA Scaled Score (z)	2,754	-0.15	0.94	2,754	-0.15	0.94	0.00
		CBT			*PBT*			
		N		*%*	*N*		*%*	
Grade Level	3	630		23	630		23	
	4	589		21	589		21	
	5	369		13	369		13	
	6	410		15	410		15	
	7	336		12	336		12	
	8	234		9	234		9	
	10	186		7	186		7	
Scale Form	r12/w23	31		1	31		1	
	r12/w12	430		16	430		16	
	r23/w23	2,277		83	2,277		83	
	r23/w12	16		1	16		1	

After matching on propensity score for these three covariates, the two groups were compared on the variables of interest: MEPA combined scaled score, reading scaled score, and writing scaled score. The results in Table 4 show that the effect sizes have now been reduced to a range of 0.11–0.16. The effect size of the MEPA combined scaled score (the one we're most concerned with) is 0.11, indicating that the CBT and PBT are comparable assessments.

Table 4. Comparison of Scaled Scores between Groups after Matching

		CBT			PBT			Effect Size
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MEPA	Combined Scaled Score	2,754	491.0	18.8	2,754	493.1	18.6	0.11
	REA Scaled Score	2,754	16.6	4.6	2,754	17.4	4.7	0.16
	WRI Scaled Score	2,754	17.0	4.4	2,754	17.5	4.1	0.12

4.2 Secondary Analysis: Additional Covariates

As described above, a secondary analysis was conducted requiring students to be matched on the additional covariates of gender, economically disadvantaged status (labeled as "EconDis" in the table; dichotomously coded as 1 if the characteristic pertained to the student, 0 otherwise), and primary language. Table 5 provides a comparison of the PBT and CBT groups in terms of all the covariates (primary and secondary) prior to matching. The two groups contain the same gender percentages, a minimal difference in percent of economically disadvantaged, and small differences in the distribution across the seven languages. Note that the sample size in Table 5 is the same as that

in Table 2, indicating that the seven languages comprise all the languages used by the students in Table 2.

Next, propensity score matching was conducted using all of the covariates. As in the primary analysis, members of the PBT group were selected to match the propensity scores of each CBT group member. Some members of the CBT group had propensity scores that could not be matched with any members of the PBT group. This resulted in the sample size being reduced from 5,508 in the primary analysis to 4,174 in the secondary analysis. Table 6 describes how well the two groups were matched on this expanded set of covariates. The results again show perfect matching.

After matching on propensity score for this expanded set of covariates, the two groups were compared in Table 7 on the variables of interest – MEPA combined scaled score, reading scaled score, and writing scaled score. The results show that the effect sizes changed only slightly, ranging from 0.12 to 0.14.

Table 5. Comparison of Expanded Covariates between Groups

		CBT			*PBT*			*Effect Size*
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MCAS 2010	ELA Scaled Score (z)	2,781	-0.15	0.95	17,432	0.02	1.01	0.18
		CBT			*PBT*			
		N		*%*	*N*		*%*	
Grade Level	3	693		21	4,552		23	
	4	689		21	4,053		20	
	5	454		14	3,025		15	
	6	489		15	2,436		12	
	7	434		13	2,166		11	
	8	310		9	1,878		9	
	10	234		7	1,750		9	
Scale Form	r12/w23	97		2	402		1	
	r12/w12	2,297		37	9,789		28	
	r23/w23	3,809		61	24,327		70	
	r23/w12	51		1	316		1	
Gender	Female	2,827		47	15,730		47	
	Male	3,180		53	17,930		53	
EconDis	Yes	4,784		84	26,987		83	
Language	Spanish	3,797		79	15,666		62	
	Portuguese	382		8	1,732		7	
	Cape Verdean	59		1	1,878		7	
	Haitian Creole	232		5	2,182		9	
	Khmer/Khmai	72		2	1,417		6	
	Vietnamese	152		3	1,213		5	
	Chinese	94		2	1,069		4	

Table 6. Comparison of Expanded Covariates between Groups after Matching

		CBT			PBT			Effect Size
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MCAS 2010	ELA Scaled Score (z)	2,087	-0.25	0.88	2,087	-0.25	0.88	0.00
		CBT			*PBT*			
		N		*%*	*N*		*%*	
Grade Level	3	484		23	484		23	
	4	453		22	453		22	
	5	287		14	287		14	
	6	291		14	291		14	
	7	254		12	254		12	
	8	183		9	183		9	
	10	135		6	135		6	
Scale Form	r12/w23	11		1	11		1	
	r12/w12	322		15	322		15	
	r23/w23	1,751		84	1,751		84	
	r23/w12	3		0	3		0	
Gender	Female	982		47	982		47	
	Male	1,105		53	1,105		53	
EconDis	Yes	1,938		93	1,938		93	
Language	Spanish	1,740		83	1,740		83	
	Portuguese	141		7	141		7	
	Cape Verdean	32		2	32		2	
	Haitian Creole	68		3	68		3	
	Khmer/Khmai	21		1	21		1	
	Vietnamese	61		3	61		3	
	Chinese	24		1	24		1	

Table 7. Comparison of Scaled Scores between Groups after Matching

		CBT			PBT			Effect Size
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MEPA	Combined Scaled Score	2,087	489.19	18.24	2,087	491.67	17.92	0.14
	REA Scaled Score	2,087	16.20	4.39	2,087	16.82	4.60	0.14
	WRI Scaled Score	2,087	16.70	4.27	2,087	17.20	3.89	0.12

4.3 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 in cases where the MCAS ELA score was not available for a student. This process yielded an additional 1,161 CBT students and an additional 6,106 PBT students to be included in the secondary analysis. Table 8 provides

descriptive statistics on the covariates of 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.

Table 8. Comparison of Covariates between Groups with Extended Score Matching

		CBT			*PBT*			*Effect Size*
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MCAS 2010	ELA Scaled Score (z)	2,781	-0.15	0.95	17,432	0.02	1.01	0.18
MEPA Fall	Scaled Score (z)	1,161	-0.24	0.99	6,106	0.05	0.99	0.29
		CBT			*PBT*			
		N		*%*	*N*		*%*	
Grade Level	3	693		21	4,552		23	
	4	689		21	4,053		20	
	5	454		14	3,025		15	
	6	489		15	2,436		12	
	7	434		13	2,166		11	
	8	310		9	1,878		9	
	10	234		7	1,750		9	
Scale Form	r12/w23	97		2	402		1	
	r12/w12	2,297		37	9,789		28	
	r23/w23	3,809		61	24,327		70	
	r23/w12	51		1	316		1	
Gender	Female	2,827		47	15,730		47	
	Male	3,180		53	17,930		53	
EconDis	Yes	4,784		84	26,987		83	
Language	Spanish	3,797		79	15,666		62	
	Portuguese	382		8	1,732		7	
	Cape Verdean	59		1	1,878		7	
	Haitian Creole	232		5	2,182		9	
	Khmer/Khmai	72		2	1,417		6	
	Vietnamese	152		3	1,213		5	
	Chinese	94		2	1,069		4	

Next, propensity score matching was conducted using all covariates, including the extended score covariate. 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 30 members of the CBT group who had a propensity score that could not be matched with any members of the PBT group. This resulted in a sample size of 3,218, an increase of 1,131 over the original secondary analysis. Table 9 shows the matching of the two groups and again indicates perfect matching.

Table 9. Comparison of Expanded Covariates with Extended Score after Matching

		CBT			*PBT*			*Effect Size*
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MCAS 2010	ELA Scaled Score (z)	2,087	-0.25	0.88	2,087	-0.25	0.88	0.00
MEPA Fall	Scaled Score (z)	1,131	-0.20	0.97	1,131	-0.20	0.97	0.00
		CBT			*PBT*			
		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	
Scale Form	r12/w23	11		1	11		1	
	r12/w12	322		15	322		15	
	r23/w23	1,751		84	1,751		84	
	r23/w12	3		0	3		0	
Gender	Female	1,386		48	1,386		48	
	Male	1,494		52	1,494		52	
EconDis	Yes	2,704		94	2,704		94	
Language	Spanish	2,305		80	2,305		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	

After matching on propensity score for the expanded set of covariates with the extended score covariate, the two groups were compared on the variables of interest – MEPA combined scaled score, reading scaled score, and writing scaled score. The results shown in Table 10 indicate that the effect sizes were slightly reduced and now ranged from 0.09 to 0.12.

Table 10. Comparison of Groups Using Expanded Covariates with Extended Score

		CBT			*PBT*			*Effect Size*
		N	*Mean*	*S.D.*	*N*	*Mean*	*S.D.*	
MEPA Spring	Combined Scaled Score	3,218	481.68	24.76	3,218	483.77	24.40	0.09
	REA Scaled Score	3,218	14.98	5.21	3,218	15.60	5.30	0.12
	WRI Scaled Score	3,218	15.28	5.25	3,218	15.88	4.95	0.12

5. SUMMARY

The MEPA program has now completed two years of testing both computer-based and paper-based tests. As part of this process a study has been conducted to evaluate the comparability of the results from these two tests. 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 for the two groups. Three effect sizes, based on three standard reported MEPA scores, were calculated as follows: 0.11 for the MEPA combined scaled scores, 0.16 for the reading scaled scores, and 0.12 for the writing scaled scores.

To verify the validity of these results, two follow-up analyses were conducted using an expanded list of covariates. Both of these analyses produced effect sizes that were similar to the effect sizes from the original analysis. The most extensive analysis reported an effect size of 0.09 for MEPA combined scaled scores and 0.12 for each of the reading and writing scaled scores.

These effect sizes fall below the threshold of 0.2 advised by the TAC, and thus do not warrant treating the CBT and PBT scores differently. Therefore, we recommend that no equating of the CBT and PBT scores occur for the spring 2011 MEPA administration.

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.

Measured Progress Department of Psychometrics and Research. (2011). *Massachusetts English Proficiency Assessment 2011 technical report*. Available from www.mcasservicecenter.com/McasDefault.asp?ProgramID=14.

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.