MICHIGAN ENGLISH LANGUAGE PROFICIENCY ASSESSMENT (ELPA)

ADMINISTRATION MODE COMPARIBILITY 2011 ADMINISTRATION

GRADE 3 THROUGH GRADE 12

Michigan Department of Education
Bureau of Assessment and Accountability

Questar Inc.

TABLE OF CONTENTS

APPENDICES ..... **ii**
FIGURES AND TABLES ..... **iii**
OVERVIEW **1**
SECTION 1. INTRODUCTION **2**
1.1 Michigan English Language Proficiency Assessment (MI-ELPA) 2
1.2 Rationale and Purpose 2

SECTION 2. PROPENSITY SCORES AND MATCHED SAMPLES **4**
2.1 Propensity Score 4
2.2 Propensity Score Matching Procedures 4
(1) Choice of Covariates 5
(2) Dealing with Missing Data on the Covariates 5
(3) Matching Methods 6
(4) Assessing the Matchign Quality 6
(5) Considering Possible Violation of Ignorable Treatment Assignment 8

SECTION 3. COMPARABILITY ANALYSES **9**
3.1 Multigroup Confirmative Factor Analysis (MGCFA) using MPLUS 9
3.2 TCCs and TIF Curves Comparison 11
3.3 Item Level Comparison 27
(1) Anchor Set and Un-Anchor Set Comparisont 27
(2) Robust Z Statistic 31
(3) Differential Item Functioning (DIF) 31
(4) Distribution Homogeneity Test and Item Mean Comparison 37
3.4 Brief Summary 37

SECTION 4. CONCLUSIONS AND DISCUSSION **38**
4.1 Conclusions 38
4.2 Discussion 38

REFERENCES **40**

APPENDICES

APPENDIX A: DIFFERENCES IN COVARIATES' MEANS AND PROPORTIONS BEFORE AND AFTER MATCHING 43

A.1: Grade 3 43
A.2: Grade 4 45
A.3: Grade 5 47
A.4: Grade 6 49
A.5: Grade 7 51
A.6: Grade 8 53
A.7: Grade 9 55
A.8: Grade 10 57
A.9: Grade 11 58
A.10: Grade 12 60

FIGURES AND TABLES

Table 2.1: N counts for CBT and PPT per Grade 6
Table 2.2: Descriptive Statistics of Logit Estimated Propensity Scores bewteen Two Modes before Matching ... 7
Table 2.3: Descriptive Statistics of Differences on Logit Estimated Propensity Scores between Each Matched Pair in the Final Data Set for Comparability Study 8
Table 3.1: Testing for Measurement Invariance with Categorical Data 9
Table 3.2.a: Testing for Measurement Invariance at Level III 10
Table 3.2.b: Testing for Measurement Invariance at Level IV 10
Table 3.2.c: Testing for Measurement Invariance at Level V 11
Figure 3.1.a: TCCs at Level III 13
Figure 3.1.b: TCCs at Level IV 14
Figure 3.1.c: TCCs at Level V 15
Figure 3.2.a: Differences between TCCs at Level III 16
Figure 3.2.b: Differences between TCCs at Level IV 17
Figure 3.2.c: Differences between TCCs at Level V 18
Figure 3.3.a: TIF Curves at Level III 19
Figure 3.3.b: TIF Curves at Level IV 20
Figure 3.3.c: TIF Curves at Level V 21
Figure 3.4.a: Differences between TIF Curves at Level III 22
Figure 3.4.b: Differences between TIF Curves at Level IV 23
Figure 3.4.c: Differences between TIF Curves at Level V 24
Table 3.3: Raw Score Corresponding to Proficient Cut for CBT and PPT for Grades 3-8 (Using Matched Students) 25
Table 3.4: N Counts of Students Who Met CBT Proficient Cut but Not the PPT Proficient Cut (From Matched Students) 25
Table 3.5: Raw Score Corresponding to Proficient Cut for CBT and PPT for Grades 3-12 (Using All Students) 26
Table 3.6: N Counts of Students Who Met CBT Proficient Cut but Not the PPT Proficient Cut (Using All Students) 26
Table 3.7: Items Excluded from the Anchor Set for CBT and PPT 27
Figure 3.5.a: Scatter Plot of All Unanchored Items' Item Difficulty Parameter Estimation across Modes—Level III 28
Figure 3.5.b: Scatter Plot of All Unanchored Items' Item Difficulty Parameter Estimation across Modes—Level IV 29
Figure 3.5.c: Scatter Plot of All Unanchored Items' Item Difficulty Parameter Estimation across Modes—Level V 30
Table 3.8: DIF Classification for MC Items 33
Table 3.9: DIF Classification for Polytomous CR Items 36
Table 3.10: DIF Analysis Results 36
Table 3.11: Number of Items Showing Significant Differences (at α=0.05) across Modes 37

OVERVIEW

This Michigan English Language Proficiency Assessment (MI-ELPA) Administration Mode Comparability for the 2011 administration is organized around four major sections—Introduction, Propensity Scores and Matched Samples, Comparability Analyses and Results, and Conclusions and Discussion. A brief overview of this report is provided below.

Section 1: Introduction

This section briefly presents the need and rationale for the comparability study reported here.

Section 2: Propensity Scores and Matched Samples

This section describes the procedures used to get matched samples for comparability.

Section 3: Comparability Analyses and Results

This section provides a description of the methods used for comparability analyses and reports the corresponding results.

Section 4: Conclusions and Discussion

This section briefly summarizes the findings of Section 3, and further discusses the issue of possible violation of ignorable treatment assignment assumption involved in propensity score estimation.

SECTION 1. INTRODUCTION

1.1 Michigan English Language Proficiency Assessment (MI-ELPA)

Details about MI-ELPA, such as test design and development, test reliability and validity, can be found in each year's technical report posted at http://www.michigan.gov/mde/0,4615,7-140-22709_31168-172665--,00.html[1].

From its first development up until Spring 2010, MI-ELPA had been administered in only one mode: paper-pencil testing (PPT). Due to the trend of computer based testing (CBT), starting from Spring 2011, Michigan schools have been encouraged to participate in the CBT of MI-ELPA. Details about the CBT administration and survey results with regard to the CBT administration in 2011 have been internally documented[2], and will not be repeated here.

Nevertheless, three pieces of information worth being mentioned here:

- The MI-ELPA CBT only used form 1 and was a linear, fixed-form CBT test. All operational items (i.e., items that contribute to the overall test score) were thus the same across forms and across modes.
- All items were presented on the screen as similar as possible to their appearance on paper, with exceptions of long reading passages where a scrolling bar was used. The CBT system also made it possible for students to review items and change their responses if they need to, which is similar to what they can do with the PPT.
- Only those voluntary buildings equipped with computers and technologies that met the minimum requirements were eligible to administer the CBT. In order to decrease the impact of computer unfamiliarity on students' test performance, students were instructed to take a practice test ahead of time to get used to the CBT environment and to determine if they were comfortable with the CBT.

1.2 Rationale and Purpose

As mentioned above, the 2011 MI-ELPA involves two administration modes: CBT and PPT. According to the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), score comparability needs to be established whenever a test is administered in different modes.

For comparability study results to be generalizable to the population, a population representative sample is needed. However, due to the voluntary participation mentioned above for the 2011

[1] Note that only the most recent MI-ELPA technical report can be found online. Previous years' technical reports are available to interested readers by contacting BAA@michigan.gov.
[2] Again, interested readers may contact BAA@michigan.gov for more information.

MI-ELPA administration, it is very likely that the CBT participants are not population representative. Theoretically, the voluntary participation decision first happened at the district and/or school building level, and then within each volunteered building, students were given the possibility of taking the PPT if they felt uncomfortable with the CBT. Therefore, it is very likely that buildings or districts with very few English language learners (ELLs) but with quite a lot of computers with high internet speed would be more willing to sign up for the CBT. It is also likely that students who are more comfortable with computers are from families with higher Socioeconomic status (SES), and are thus more likely to be higher achievers academically. Such possible ability differences will make the comparability analysis results questionable if the comparison is done directly on the two sets of students (i.e., students who volunteered to take the CBT vs. the rest of the student population who took the PPT).

In order to form comparable samples for comparison, propensity score matching was used to select a group of students from the PPT group that are comparable to the CBT group. This technique is described in detail in Section 2. Mode comparison analyses were then carried out using the matched samples[3], and related details can be found in Section 3. Due to the fact that the comparison was carried out on matched samples, which were found to be not population representative, the main goal of this report was to confirm the decision made at the scoring stage that the CBT students can be treated as if they took the PPT and the item calibration and scaling procedures being carried out using combined data was appropriate.

[3] Only one analysis—real data impact overall in classification inconsistency—was carried out using all data, as can be seen in Section 3.

SECTION 2. PROPENSITY SCORES AND MATCHED SAMPLES

This section describes how matched samples are formed. Specifically, the concept of propensity score was introduced here, along with the description of propensity score matching procedures.

2.1 Propensity Scores

A propensity score, which does not depend on the response information, is the conditional probability of assignment to treatment (in current report, take CBT instead of PPT) given the covariates (Rosenbaum & Rubin, 1985). With a dichotomous treatment variable, logistic regressions with the treatment assignment as an outcome are used to estimate the propensity scores (e.g., Harder, Stuart, & Anthony, 2010). In this report, the CBT was coded as 1 and the PPT was coded as 0. After propensity scores are estimated, different approaches such as matching, weighting, and subclassification can then be applied to form comparable groups (Harder et al., 2010; Stuart, 2010).

In this report, we only considered matching, and in particular, the pair matching in forming comparable groups with the same sample size in order to get rid of 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: (1) choice of covariates, (2) dealing with missing data on the covariates, (3) matching methods, and (4) assessing the matching quality (Lottridge et al., 2011; Steiner, Cook, Shadish, & Clark, 2010; Stuart, 2010) . Note, however, unlike randomization which can balance both observed and unobserved covariates, matching on propensity scores can only balance on observed covariates (Rosenbaum & Rubin, 1985). Therefore, the fifth issue to consider when conducting propensity score matching is the possible violation of ignorable treatment assignment after analyzing the outcome of interest. All of these issues are discussed in more detail later, and the fifth one is also tackled with in the Conclusion and Discussion section.

Since propensity score matching approach can only balance observed covariates, the resulting "equivalent" ability groups is an approximation. Lottridge et al. (2011) found that with statewide high school end-of-course tests in Algebra and English, the propensity score matching approach produced similar results to the single group with counterbalancing design (which was considered a credible benchmark method by Lottridge et al. [2011]), using computer skill scores, student level demographics variables, and previous achievement scores as covariates in estimating the propensity score of taking CBT. Even though the results cannot be readily generalized to all mode comparison studies, the findings are encouraging because it shows that the propensity score matching technique is promising when the single group with counterbalancing design is not feasible for conducting mode comparison studies.

2.2 Propensity Score Matching Procedures

This section provides detailed information on the five issues mentioned above when conducting propensity score matching.

(1) Choice of Covariates

Three types of covariates can be included in a propensity score estimation model:

- a variable related to both the outcome and the treatment,
- a variable related to only the outcome, and
- the variable related only to the treatment.

Simulation studies found that the optimal propensity score model was the one only included the first two types of variables mentioned above (i.e., a variable related to both the outcome and the treatment, and a variable related to only the outcome) (Brookhart et al., 2006). Steiner et al. (2010) found that the first type mentioned above (i.e., a variable related to both the outcome and the treatment) was crucial for removing bias. Pre-test measures were found to be highly correlated to potential outcomes (Steiner et al., 2010), and were therefore suggested to be included as covariates for estimating propensity scores. Socioeconomic status (SES) is a student-level variable that is possible to be relevant to any score differences across mode (Pomplun, Ritchie, & Custer, 2006). Way, Lin, & Kong (2008) used all possible prior achievement scores in their mode comparability study. In order to fully utilize the capability of propensity scores in balancing multiple covariates, we included all possible recent previous achievement scores (including the previous year's MI-ELPA score), all available demographic variables at the student level, and also the school-level background variables such as the number of ELLs, the number of computers with high-speed internet per ELL, school location, and achievement variables (created as the mean of the school-level aggregates of the previous three years). Since different general assessments were administered for different grade levels, the previous achievement scores that were included in each grade level's propensity score estimation were 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 at hand, more than half of the schools had very few (≤ 5) ELLs, so a multilevel method was not practical for this data. In order to still taking school level covariates into consideration in estimating the propensity scores, the school level covariates were used as individual level covariates.

(2) Dealing with missing data on the covariates

Since very few students took the CBT in 2011 (see Table 2.1 for detailed N counts information for different testing mode per grade), we used imputation to keep all students in the propensity score estimation model. Although various missing data handling techniques have been proposed in the context of propensity score estimation, no significant differences in treatment effect estimations were found between various techniques applied to real data sets (Harder et al., 2010). For this report, a multiple imputation procedure 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 multiple imputed values, we only used the mode or

median of the five sets of imputed values to simplify the analysis[4]. Since covariates are from both the school-level and the student-level, we conducted multiple imputations in two steps. First, multiple imputations were conducted at the school level, and the mode or median of the five imputed values were computed. Then the imputed school level data were combined with student level variables to conduct multiple imputations at the student level. Again, mode or median of the five sets of imputed values at the students level were used to form the final single data set to conduct comparison analysis.

Table 2.1. N Counts for CBT and PPT per Grade.

Grade	CBT	PPT
3	308	6,603
4	243	5,774
5	271	4,772
6	144	3,902
7	121	3,721
8	136	3,559
9	196	4,010
10	152	3,532
11	139	2,593
12	115	2,072

(3) Matching Methods

Different matching methods exist in the literature, such as the nearest neighbor matching approach (Stuart, 2010). The optimal matching algorithm is found to be better than the nearest neighbor matching approach (or the greedy algorithm) for pair matching with a large pool of controls (Hansen, 2004). As shown in Table 2.1, there is a huge pool of PPT students in comparison to the CBT students at each grade level. Therefore, the optimal matching algorithm was used. Specifically, the R package OPTMATCH (Hansen & Klopfer, 2006) was used to conduct pair matching based on the logit estimated propensity scores per grade level.

(4) 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, two-sample t-tests before and after matching, and the comparison of pseudo-R^2 before and after matching (Caliendo & Kopeinig, 2008). In this report, the balance check approach used in the R package RItools (Bowers, Fredrickson, & Hansen, 2010) was chosen 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 &

[4] In essence, it is really a single imputation approach then, as only one set of imputed values was used for later analysis. Median was only used when mode was not found.

Bowers, 2008). If the balance test showed significant results, the whole procedure from propensity score estimation would have needed to be repeated. In order to obtain groups as similar as possible for comparison, we also excluded matched pairs if their differences of logit estimated propensity scores were outside of the range of [Q1-1.5*interquartile range, Q3+1.5*interquartile range]. Here Q1 and Q3 represent first and third quartiles of the differences respectively.

Even though the omnibus measure of balance was used in judging matching quality in this study, we still presented comparison results before and after matching for each covariate as additional information (see Appendices A.1 to A.10). For these tables, the χ^2 tests were used for dichotomously coded variables and t-tests were used for continuous variables. Because our purpose of propensity score matching was to form comparable groups with regard to English language ability, only achievement-related variables at the student and school levels, student demographic variables, the number of ELLs at the school level (per grade), and the number of computers with high-speed internet per ELL at the school level (per grade) were presented in these tables. In addition, descriptive statistics of logit estimated propensity scores for the two groups before matching were presented (see Table 2.2), along with the descriptive statistics of the differences of logit estimated propensity scores between each matched pair (see Table 2.3) in the final data set (i.e., the data set for the comparability analysis).

Table 2.2. Descriptive Statistics of Logit Estimated Propensity Scores between Two Modes before Matching

	N	*Mean*	*SD*	*Minimum*	*Maximum*
Grade 3					
CBT	308	-1.947	1.175	-12.505	0.484
PBT	6,603	-7.348	6.780	-29.586	0.476
Grade 4					
CBT	243	-1.566	1.496	-7.688	2.221
PBT	5,774	-7.748	6.960	-29.529	4.203
Grade 5					
CBT	271	-1.400	1.260	-8.067	1.232
PBT	4,772	-7.918	7.403	-28.335	5.144
Grade 6					
CBT	144	-1.379	1.539	-4.925	2.321
PBT	3,902	-8.943	7.625	-29.338	2.900
Grade 7					
CBT	121	-1.334	1.905	-6.103	2.356
PBT	3,721	-6.081	2.980	-26.124	2.782
Grade 8					
CBT	136	-1.283	1.441	-4.654	1.051
PBT	3,559	-6.984	4.320	-29.398	4.001
Grade 9					
CBT	196	-1.465	1.705	-6.209	3.743
PBT	4,010	-5.742	3.806	-28.113	2.796
Grade 10					
CBT	152	-1.708	1.250	-5.100	1.952

PBT	3,532	-6.598	4.614	-20.749	2.814
Grade 11					
CBT	139	-0.920	1.762	-5.969	3.051
PBT	2,593	-5.961	3.785	-29.610	12.410
Grade 12					
CBT	115	-1.075	1.963	-6.401	3.113
PBT	2,072	-5.928	3.804	-29.046	1.751

Table 2.3. Descriptive Statistics of Differences on Logit Estimated Propensity Scores between Each Matched Pair in the Final Data Set for Comparability Study

Grade	*N*	*Mean*	*SD*	*Minimum*	*Maximum*
3	285	0.000	0.002	-0.005	0.006
4	190	0.000	0.003	-0.008	0.011
5	211	0.000	0.003	-0.008	0.008
6	112	0.001	0.005	-0.015	0.020
7	100	0.046	0.085	-0.005	0.370
8	109	0.015	0.032	-0.084	0.147
9	171	-0.028	0.060	-0.228	0.021
10	124	0.000	0.003	-0.009	0.009
11	119	-0.129	0.199	-0.896	0.011
12	106	-0.262	0.426	-1.535	0.011

Based on these tables, we see that the matching was quite successful per grade level. Grade 11 and Grade 12 seems to have been matched less well in comparison to other grade levels, and this may be due to the relatively smaller PPT pool to get matched samples from, and also the fact that grade 11 and grade 12 have more variables in consideration for propensity score estimation, thus involves the use of more imputed values.

(5) Considering Possible Violation of Ignorable Treatment Assignment

As stated by Rosenbaum and Rubin (1985), propensity score approaches cannot balance unobserved variables. Therefore, if an unobserved variable is significantly related to both the treatment assignment and the outcome but is unmeasured and thus is not included in the propensity score estimation model, the resulting treatment effect estimates would be biased (Stuart, 2010). Different sensitivity analysis approaches have been proposed in the literature (Caliendo & Kopeinig, 2008; Stuart, 2010). According to Rosenbaum (2010), such analyses are conducted by altering the chances of receiving treatment for those units that appear to have similar chances. The examples included in Rosenbaum (2010) all indicated a significant effect being found, and the sensitivity analyses were trying to specify when such an effect becomes non-significant statistically. In our case, however, we hoped to conclude that the two modes are comparable (i.e., that no significant differences can be found between the two modes). Therefore, if a sensitivity analysis is conducted, the direction will be the opposite (i.e., trying to specify when the two modes will show statistically significant differences). No such sensitivity analyses could be found so far based on the authors' knowledge. However, we still discuss the possibility of bias in the Conclusions and Discussion section.

SECTION 3. COMPARABILITY ANALYSES AND RESULTS

This section describes the methods and results for comparability analyses based on the matched samples obtained per grade level.

Note that even though 10 different grade levels are involved here, only 3 different assessments (if only counting the operational items) are under investigation. Specifically, grades 3 to 5 ELLs were administered the Level III MI-ELPA, grades 6 to 8 ELLs were administered the Level IV MI-ELPA, and grades 9 to 12 were administered the Level V MI-ELPA. Therefore, for analyses involved in this section, students from corresponding grade levels were combined for each mode for each assessment level.

Three sets of comparison analyses were conducted, with the first two focusing on the overall test, and the last one focusing on each individual item. First, multigroup confirmatory factor analysis (MGCFA) using MPLUS (Muthén & Muthén, 2009) was conducted. Second, test characteristic curves (TCC) and test information functions (TIF) were compared. Third, various item level analyses were conducted, such as anchor set and un-anchor set comparison, Robust Z statistic of item difficulty estimation, differential item functioning, and multiple choice (MC) option distribution or constructed response (CR) score distribution homogeneity test and item mean comparison. The details of the three sets of analyses and corresponding results are presented below.

3.1 Multigroup Confirmatory Factor Analysis (MGCFA) using MPLUS

For this analysis, three nested models were compared to establish measurement invariance (Schroeders & Wihelm, 2011): configural invariance, strong invariance, and strict invariance. For configural invariance, factor loadings and thresholds are freely estimated but the residual variances are fixed at 1 and factor means are fixed at 0 in both groups. For strong invariance, the factor loadings and thresholds are fixed to be equal across the two groups, the residual variances are fixed at 1 for the PPT group but are freed in the CBT group, and the factor means are fixed at 0 in the PPT group but are freed in the CBT group. The only difference between the strict invariance model and the strong invariance model is that the former also fixes the residual variances at 1 in both groups (see Table 3.1 below, which is adapted from Schroeders & Wihelm, 2011).

Table 3.1. Testing for Measurement Invariance with Categorical Data

Invariance Type	Factor Loadings	Thresholds	Residual Variances	Factor Means
Configural invariance	(*	*)	Fixed at 1	Fixed at 0
Strong invariance	(Fixed	Fixed)	Fixed at 1/*	Fixed at 0/*
Strict invariance	(Fixed	Fixed)	Fixed at 1	Fixed at 0/*

Note. From Schroeders & Wihelm (2011). The asterisk (*) indicates that the parameter is freely estimated. Fixed=the parameter dominated in the title of the column is fixed to equity across groups; Fixed at 1=the residual variances are fixed at 1 in both groups; Fixed at 1/*=the residual variance is fixed at 1 in one group whereas freed in the other group; Fixed at 0=factor means are fixed at 0 in both groups. Fixed at 0/*=the factor mean is fixed at 0 in one group and freed in the other. Parameters in parentheses need to be varied in tandem.

If strict invariance is established, the observed scores can be considered as interchangeable (Neuman & Baydoun, 1998). If, however, only the strong invariance is established, ability estimates can be considered as comparable when residual item variances can be attributed to random error (Schroeders & Wilhelm, 2011). Same as in Schroeders & Wihelm (2011), here we estimated all models using the default estimator—weighted least squares mean and variance adjusted (WLSMV) estimator with Theta parameterization. Due to problems found with the χ^2 statistics (Chen, 2007; Cheung & Rensvold, 2002), the following fit indices and cutoff criteria were used: the comparative fit index (CFI) ≥ 0.95 and the root mean square error of approximation (RMSEA) < 0.05 (Hu & Bentler, 1998) for indicating a good model fit; and a change of ≥ -.010 in CFI (Chen, 2007; Cheung & Rensvold, 2002) and a change of ≥ .015 in RMSEA (Chen, 2007) for indicating invariance for each step of the nested model comparison.

Table 3.2.a. Testing for Measurement Invariance at Level III

	CFI	RMSEA	ΔCFI	ΔRMSEA
1-factor				
Configural invariance	0.880	0.052		
Strong invariance	0.883	0.051	0.003	-0.001
Strict invariance	0.915	0.049	0.032	-0.002
2-factors				
Configural invariance	0.958	0.031		
Strong invariance	0.957	0.031	-0.001	0
Strict invariance	0.965	0.031	0.008	0

Table 3.2.b. Testing for Measurement Invariance at Level IV[5]

	CFI	RMSEA	ΔCFI	ΔRMSEA
1-factor				
Configural invariance	0.919	0.053		
Strong invariance	0.921	0.052	0.002	-0.001
Strict invariance	0.938	0.048	0.017	-0.004
2-factors				
Configural invariance	0.953	0.040		
Strong invariance	0.953	0.040	0	0
Strict invariance	0.964	0.037	0.011	0.003

[5] Note that due to data limitation at this level—one category of the last Speaking item does not contain the lowest possible value "0" for the CBT group, we have to add in two lines of fake data (the two lines are rotating possible minimum scores and possible maximum scores on items) to make the program run. Therefore, the corresponding results should be interpreted with caution.

Table 3.2.c. Testing for Measurement Invariance at Level V

	CFI	RMSEA	ΔCFI	ΔRMSEA
1-factor				
Configural invariance	0.945	0.053		
Strong invariance	0.946	0.052	0.001	-0.001
Strict invariance	0.967	0.048	0.021	-0.004
2-factors				
Configural invariance	0.976	0.035		
Strong invariance	0.976	0.035	0	0
Strict invariance	0.984	0.033	0.008	0.002

As shown in Tables 3.2.a to 3.2.c, for all assessment levels, we tried both 1-factor model and the 2-factors model. The 1-factor model was in essence the one used in the scaling procedure, as the types of Item Response Theory (IRT) models we applied to the data assumed unidimensionality. However, additional exploratory data analysis indicated that the data may better fit a 2-factor model, with Speaking items forming their own factor. As can be seen in the above tables, using the good model fit criteria, the 2-factors model would be the ones chosen for all assessment levels. Using the change of CFI criterion, strict invariance holds for all assessment levels, but when the change of CFI criterion was supplemented by the change of RMSEA criterion, then even the strong invariance does not hold. However, as mentioned by Chen (2007), "changes in RMSEA are more likely to be affected by sample size and model complexity" (p. 501), thus CFI was chosen as the main criterion. We thus still consider here that the strict invariance appears to hold.

3.2 TCCs and TIF[6] Curves Comparison

The test characteristic function (or graphically the test characteristic curve (TCC)) is "the expected raw score on the total test given θ; the expected raw score is the sum of the item scores assigned to each item response times the probability of obtaining that item response" (Yen & Fitzpatrick, 2006, p. 125):

$$E(X.|\theta) = \sum_{i=1}^{n} E(X_i|\theta) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} x_{ij} P(X_i = x_{ij}|\theta).$$

In addition, "when the optimal weights are used, the test score information equals the sum of the item information functions, which equals the test information" (Yen & Fitzpatrick, 2006, p. 127):

$$I(\theta) = \sum_{i=1}^{n} \frac{[P_i'(\theta)]^2}{P_i(\theta)[1-P_i(\theta)]} = \sum_{i=1}^{n} I(X_i|\theta).$$

Detailed mathematical formulas for the probability of obtaining the item response for each item can be found in Section 6 of the posted technical report, and the mathematical formulas for item information can be found in Section 7 of the posted technical report.

[6] TIF stands for Test Information Function.

For this analysis, TCCs and TIF curves, along with the difference between the expected raw scores and the differences between test information along the whole ability scale, were plotted to show possible differences when item calibration and scaling were carried out separately by mode (i.e., different conversion tables were used). Because the most important classification on the test is whether students are proficient or not, differences in the expected raw scores between the two groups at the proficient cuts were examined. If differences in expected raw scores are below 0.5, there will be no difference in classification of the two groups[7]. However, if differences in expected raw scores are at or above 0.5, theoretically there will be some classification differences. If such large differences are found, the real test population will then be examined to see how many students would be classified differently as a real impact investigation. If no students are found to be impacted in the real test population, even with theoretical differences, it is still considered operationally appropriate to combine the two groups when classification accuracy is the main concern.

[7] The scale scores are linear transformation of theta scales, which are shown as X-axis in Figures 3.1.a to 3.4.c. The proficient cut was originally defined on the theta scale, thus similarly also on the scale score scale. A student will be classified as proficient if he/she obtains at least the theta value (or scale score) at the proficient cut. Such classification does not change depending on how much a person's score is above the cut. What matters is only that one cut value. If the cut value corresponds to the same raw score point, then no difference will be found for student classification. However, if the cut point correspond to different raw score points for students taking CBT or PPT, then a student may be classified differently using different conversion tables. Due to rounding, when difference between two TCCs have at least 0.5 expected raw score for a particular theta cut, the cut will be associated with different raw score points, which means that it is likely that students may be classified differently based on which conversion table to use. Nevertheless, in practice, it may happen that no student obtains the exact raw scores in question. If that is the case with real data, then classification inconsistency will not occur in practice. Therefore, when theoretical differences are found, the natural next step is to examine the real data impact.

Figure 3.1.a. TCCs at Level III

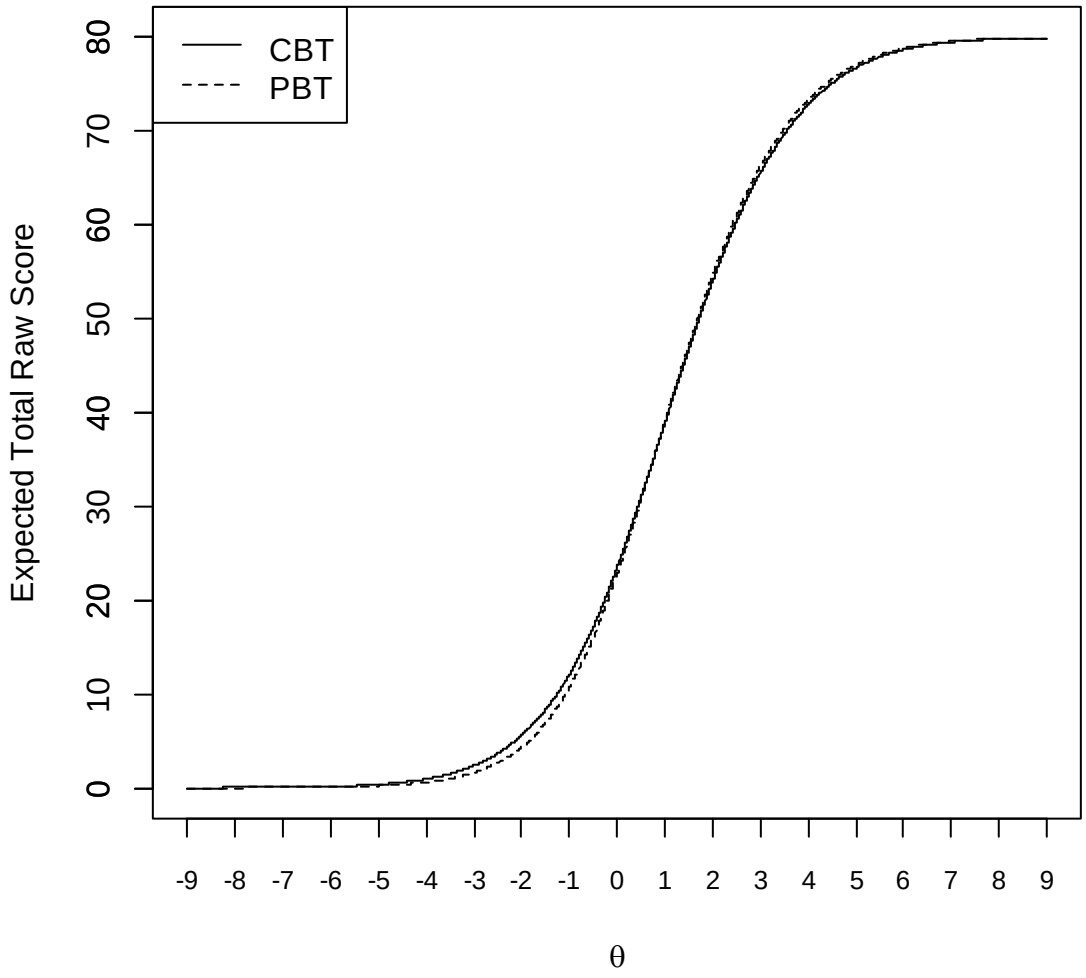


Figure 3.1.b. TCCs at Level IV

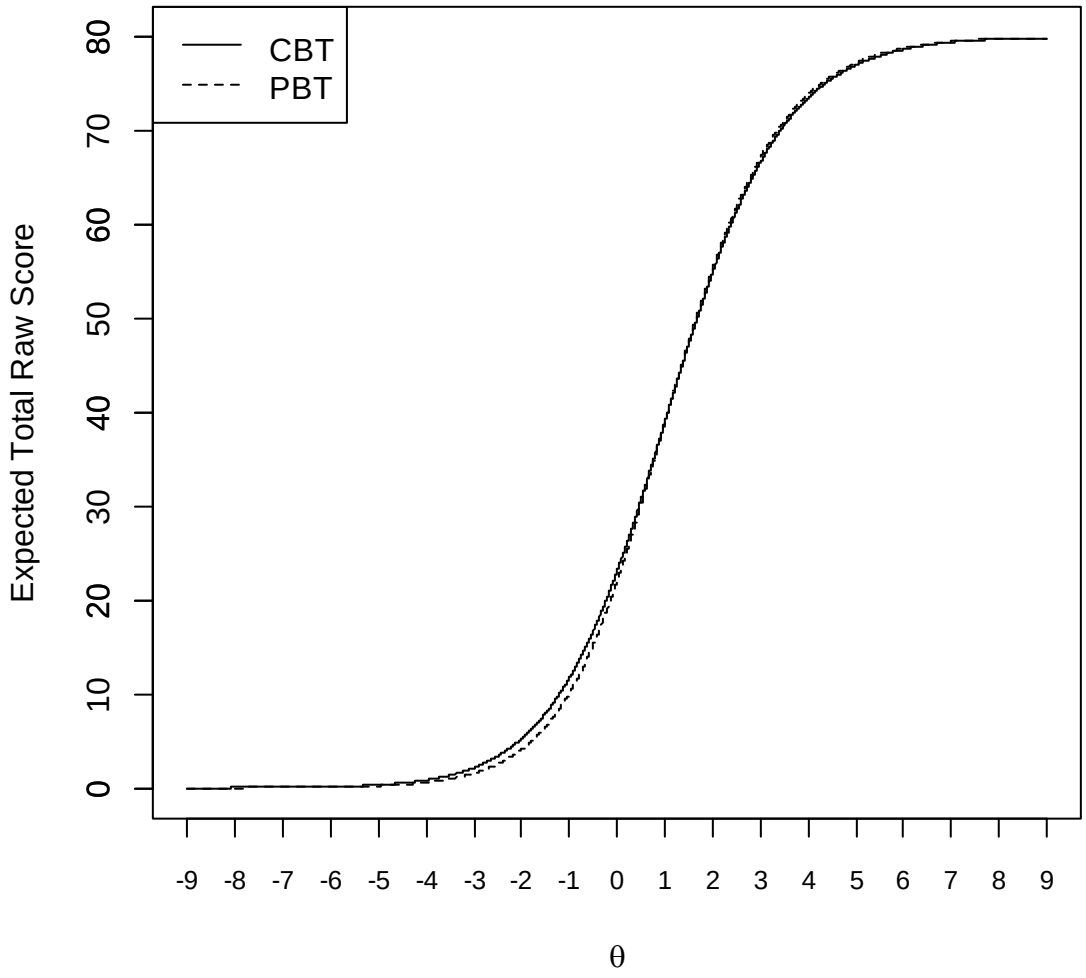


Figure 3.1.c. TCCs at Level V

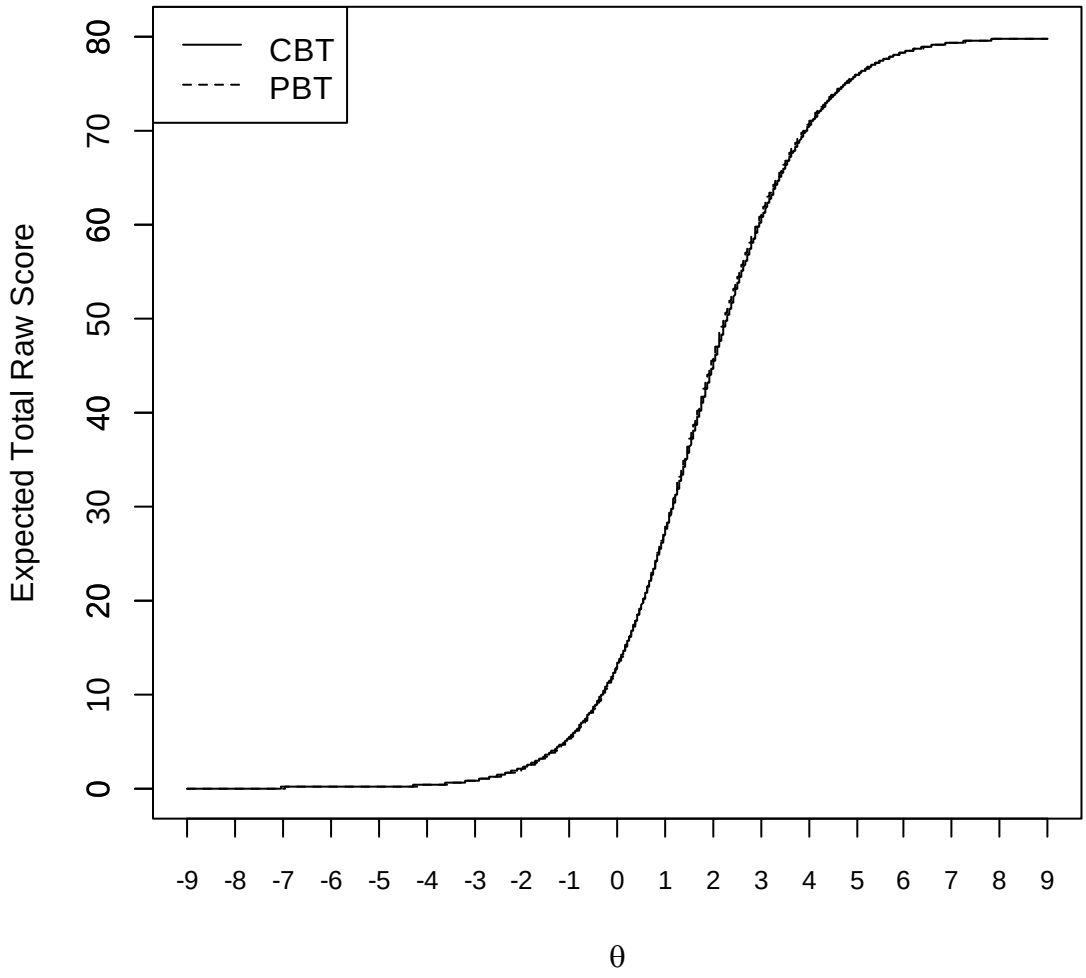


Figure 3.2.a. Differences between TCCs at Level III

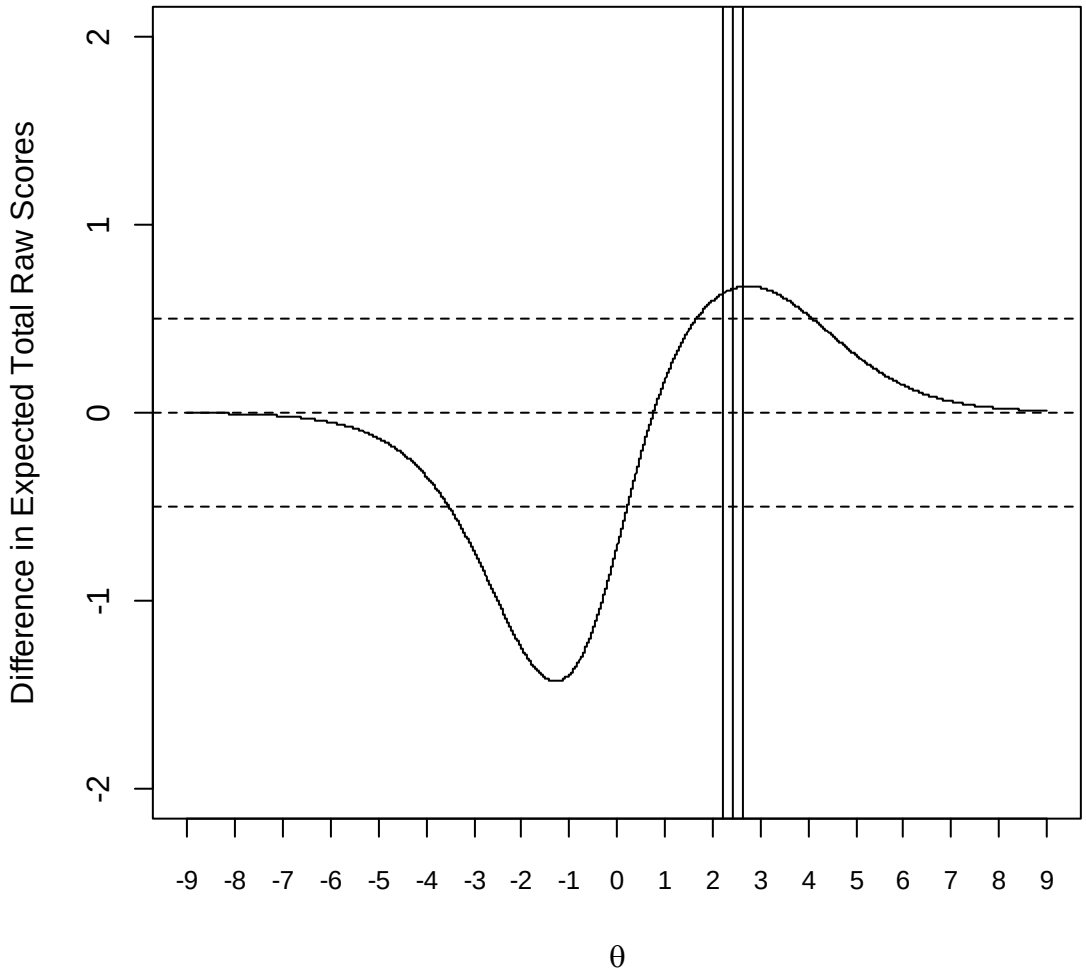


Figure 3.2.b. Differences between TCCs at Level IV

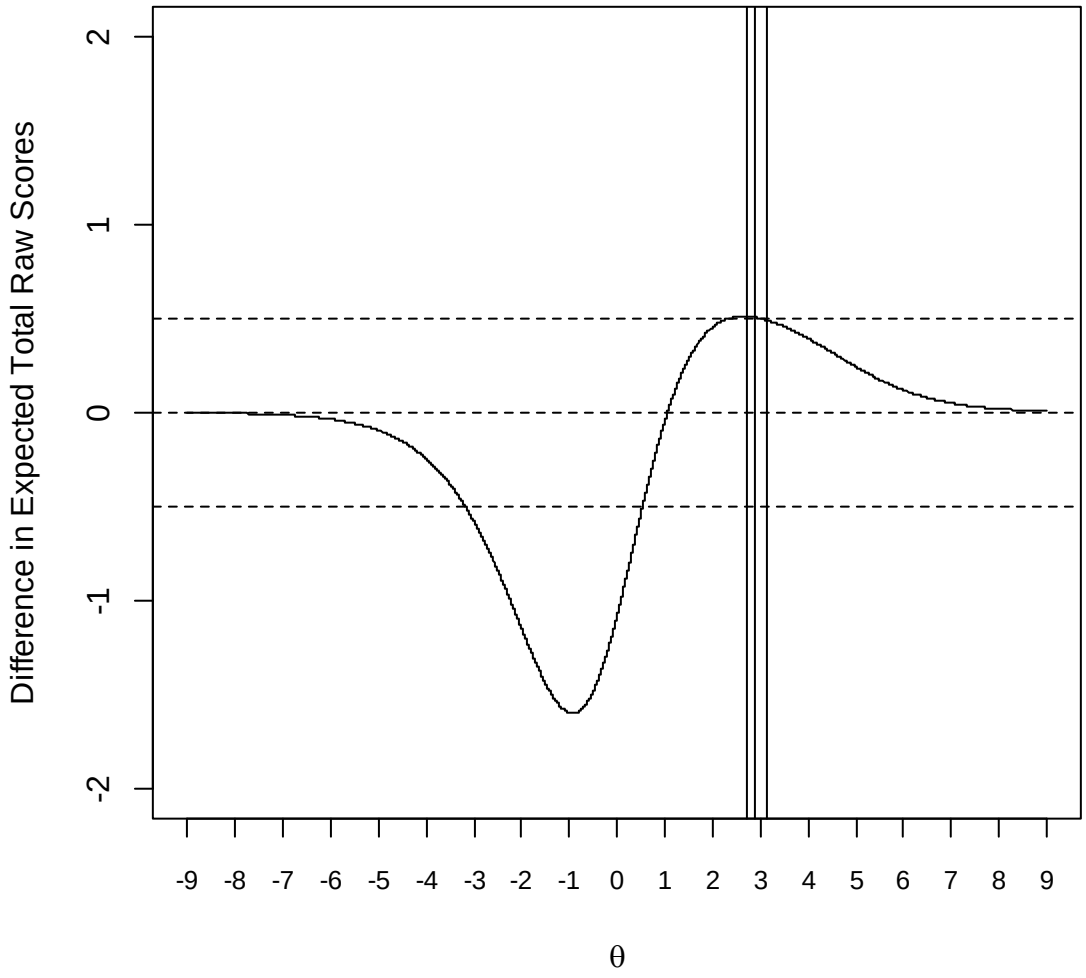


Figure 3.2.c. Differences between TCCs at Level V

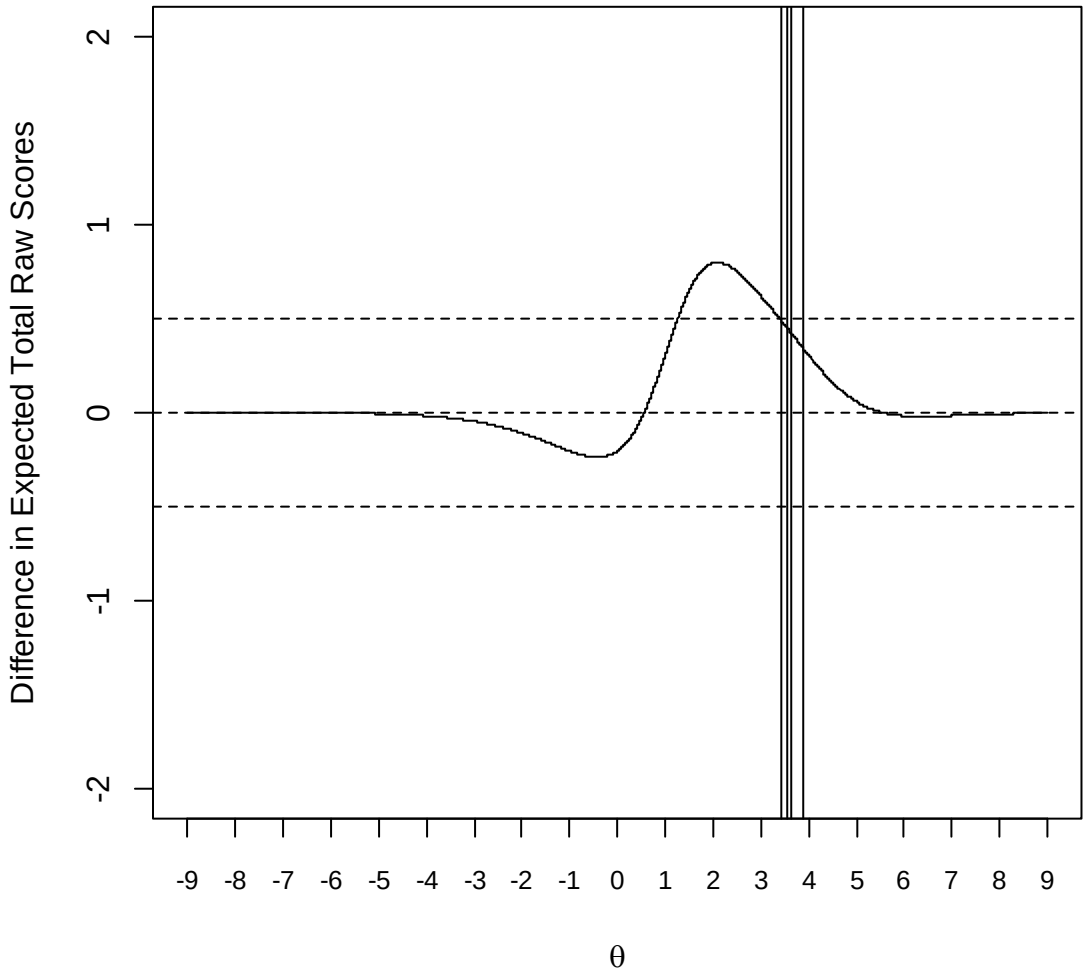


Figure 3.3.a. TIF Curves at Level III

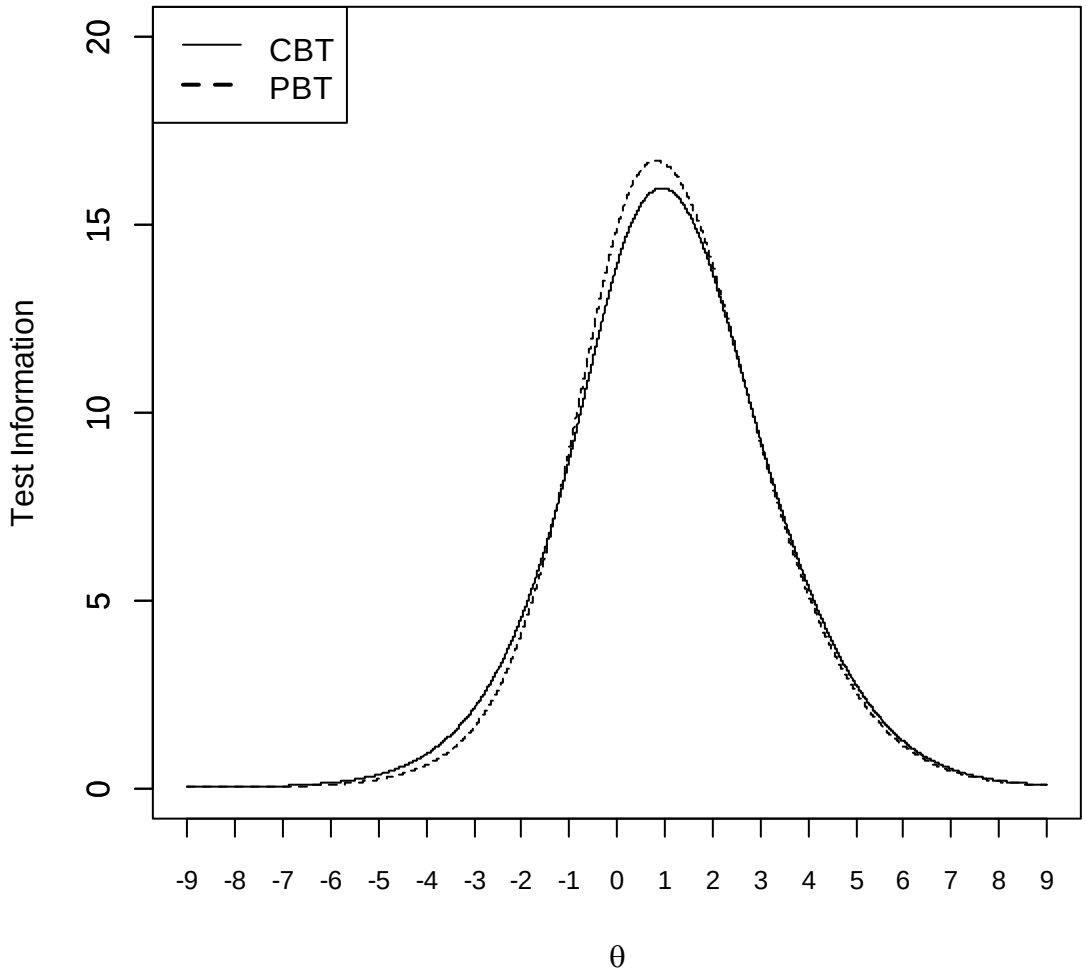


Figure 3.3.b. TIF Curves at Level IV

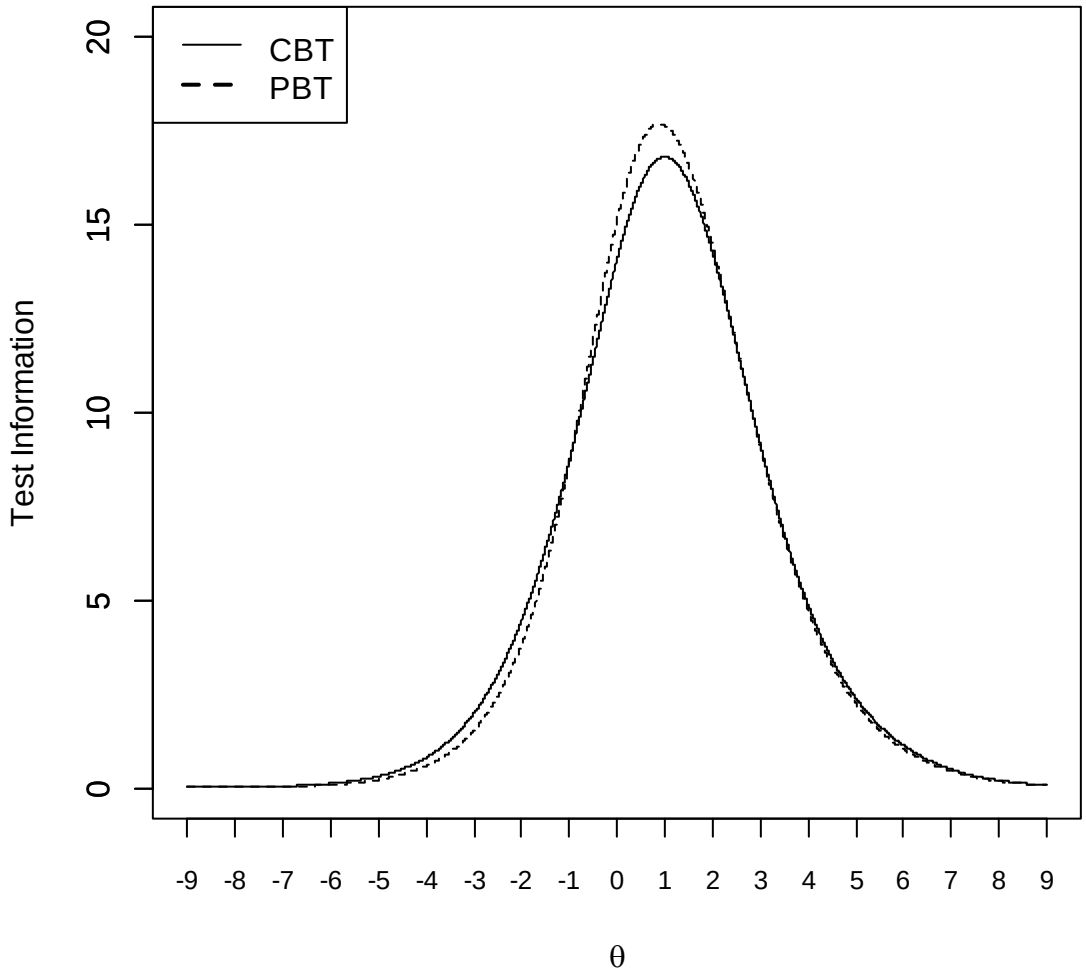


Figure 3.3.c. TIF Curves at Level V

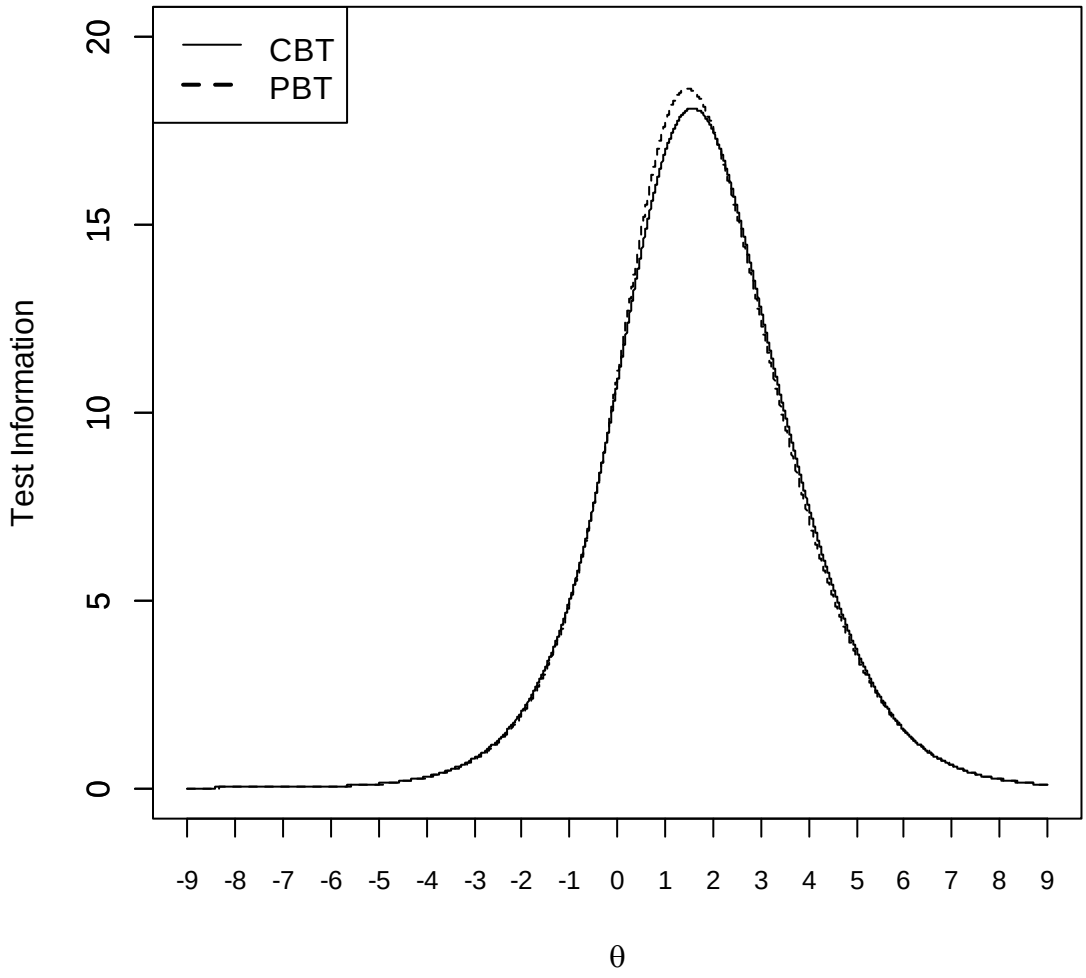


Figure 3.4.a. Differences between TIF Curves at Level III

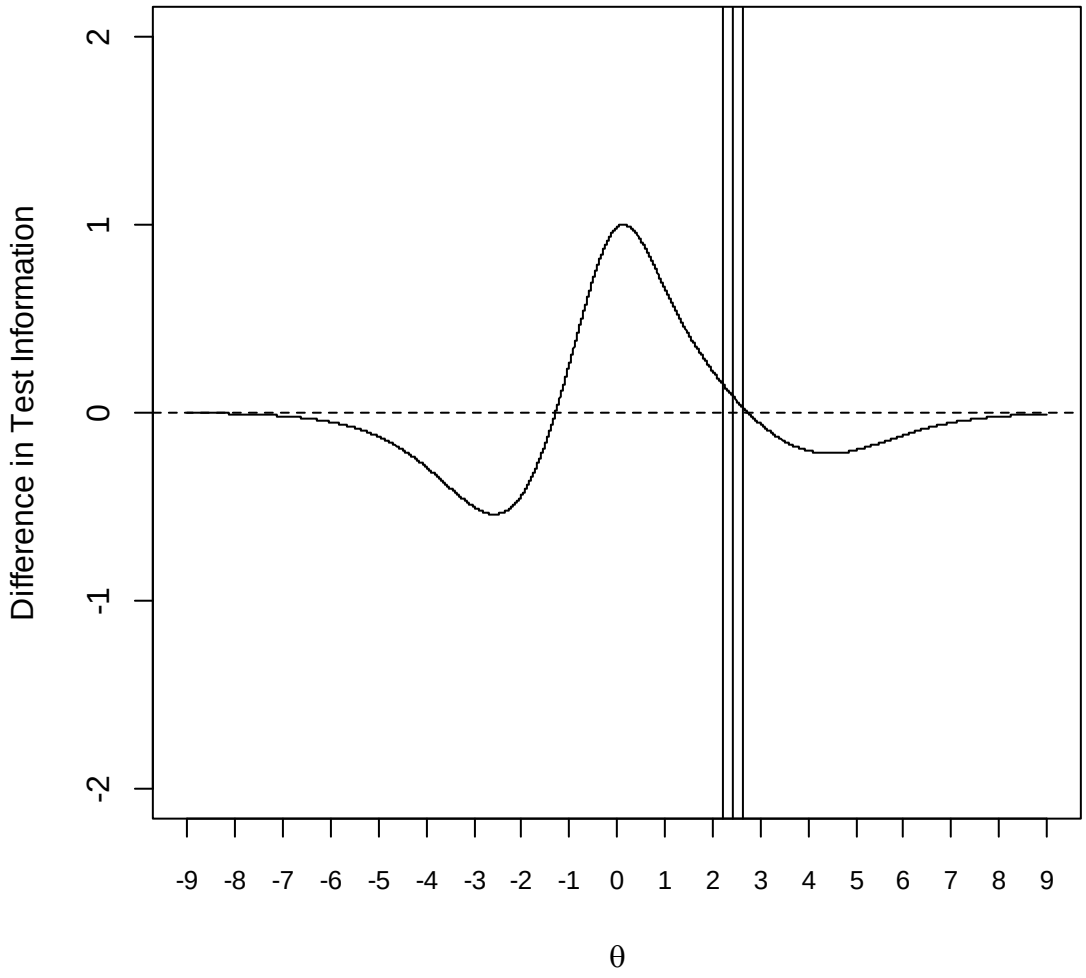


Figure 3.4.b. Differences between TIF Curves at Level IV

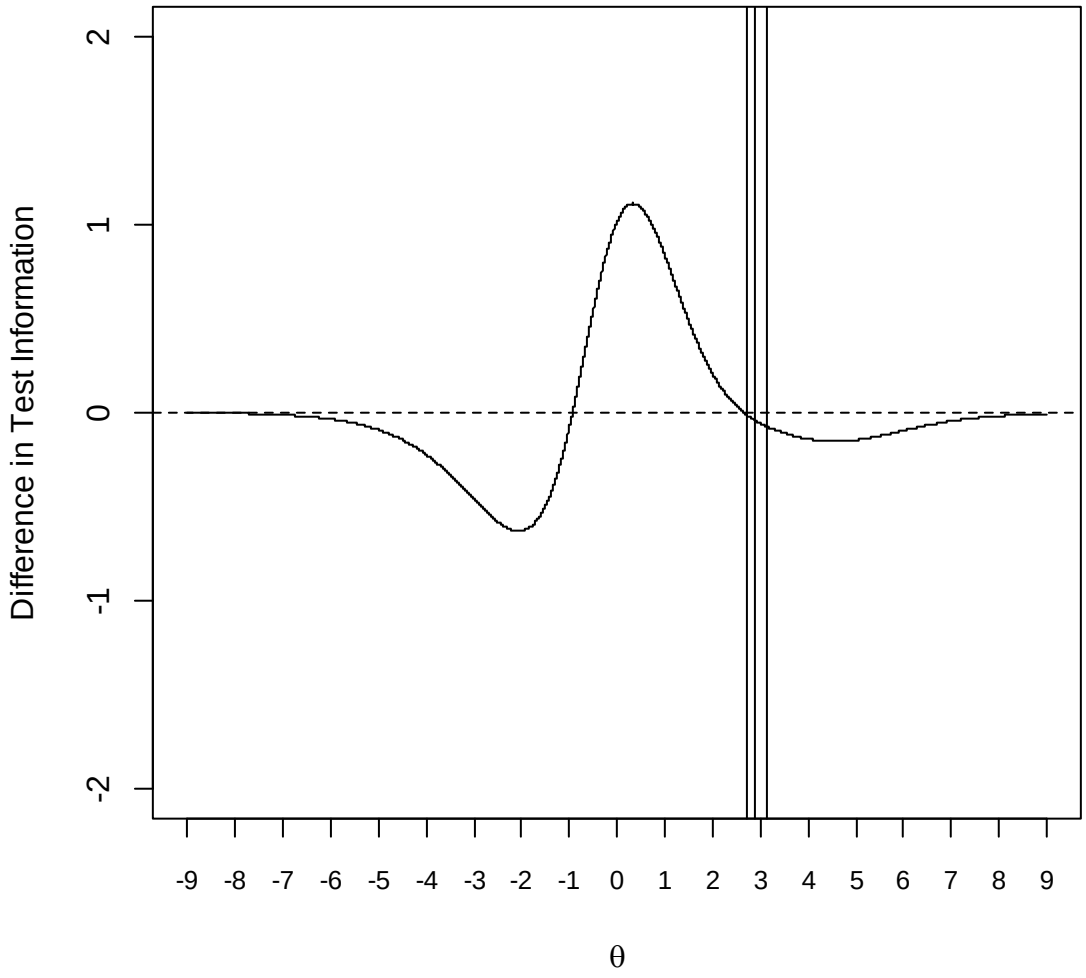
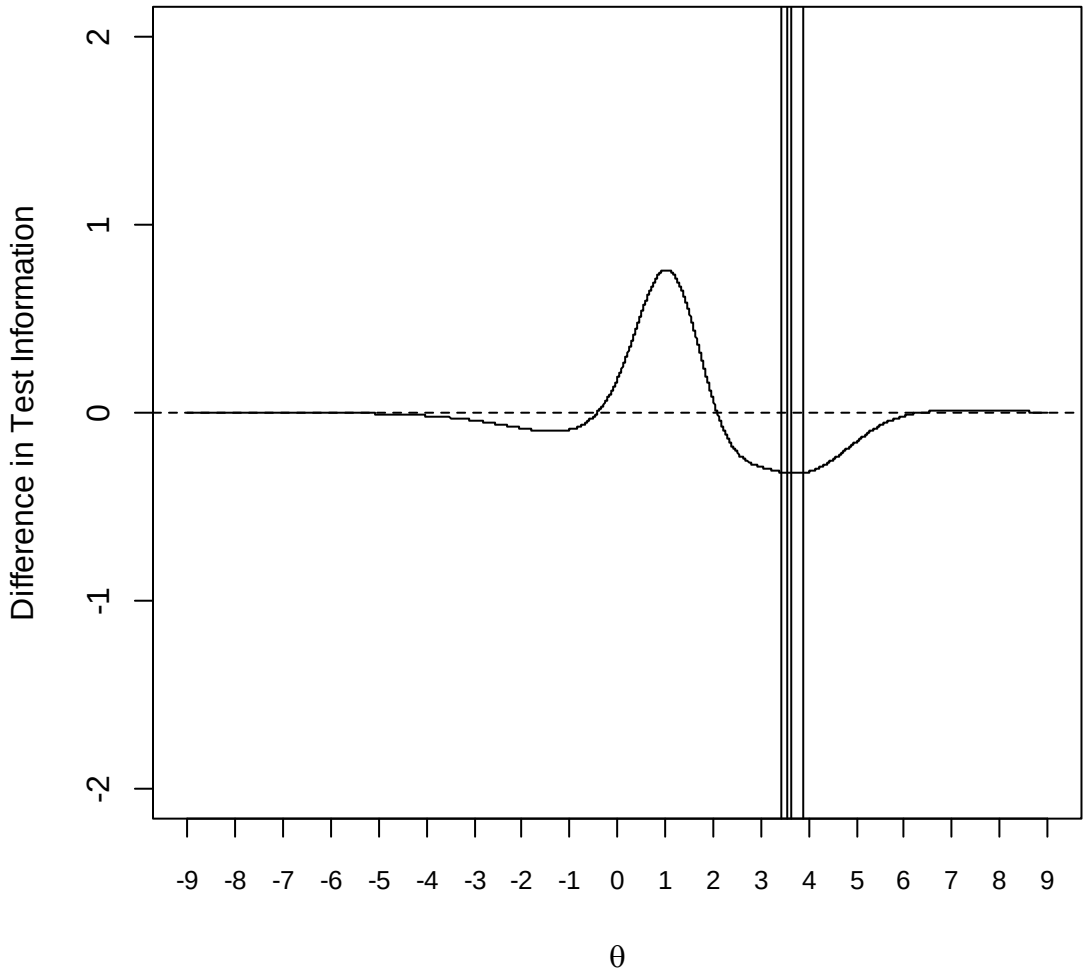


Figure 3.4.c. Differences between TIF Curves at Level V



As shown above, Figures 3.1.a to 3.1.c present TCCs at each assessment level, Figures 3.2.a to 3.2.c demonstrate differences between TCCs at each assessment level, Figures 3.3.a to 3.3.c. present TIF curves at each assessment level, and Figures 3.4.a to 3.4.c demonstrate differences between TIF curves at each assessment level. For easy reference, a discrepancy of 0.5 raw score point was included in Figures 3.2.a to 3.2.c as two horizontal dotted lines, and the vertical lines in these graphs and also in Figures 3.4.a to 3.4.c. indicate proficient cut per grade level.

All TCC graphs indicate that the TCCs for CBT and PPT are close to each other for each assessment level, especially for Level V. This is further confirmed when the differences between TCCs graphs are examined. As explained in detail in footnote 7, the focus here is on consistent classification and difference in expected raw scores less than 0.5 will not lead to inconsistent classification. Figures 3.2.a indicated that there may be classification inconsistency due to the differences of larger than 0.5 expected raw scores at all three proficient cuts in that graph.

Figure 3.2.b seems to indicate borderline scenario, as the differences seem to be at around 0.5 for those three proficient cuts at Level IV. Figure 3.2.c clearly indicates that the differences are less than 0.5 expected raw scores at all four proficient cuts. Therefore, for Level III and Level IV students, real data impact needs to be further examined to see how many students are classified differently using two different conversion tables. Figures 3.3.a to 3.4.c basically indicate that one mode of test delivery may provide more information than the other in different ability (i.e., theta) ranges, and there is no consistent advantage in providing more information (i.e., accuracy of estimation) of one mode over the other.

Note that since the focus here is whether CBT students can be treated as if they were PPT students, we only examined the classification impact on CBT students if different conversion tables were applied.

Based on the IRT calibration results using the matched samples, we found that the following raw score points correspond to the application of proficiency cut on the theta scale (or equivalently, on the scale score scale) for each grade from grade 3 to 8 (i.e., Levels III and IV) (see Table 3.3).

Table 3.3. Raw Score Corresponding to Proficiency Cut for CBT and PPT for Grades 3-8 (Using Matched Students).

Grade	CBT	PPT
3	57	58
4	60	61
5	62	63
6	64	65
7	66	67
8	68	69

As shown in Table 3.3, students who obtain raw score 57 at grade 3 are classified as proficient if using the CBT conversion table, but are classified as NOT proficient if using the PPT conversion table. However, if no real student obtain score point 57, then there should be no real data impact for grade 3 CBT students. The real data impact for CBT students is presented in Table 3.4.

Table 3.4. N Counts of Students Who Met CBT Proficient Cut but not the PPT Proficient Cut (From Matched Students).

Grade	N Counts
3	10
4	8
5	8
6	4
7	5
8	7

It can be argued that for such classification impact, it is more practical to use all CBT students and all PPT students, instead of just the matched samples. Different calibrations were thus carried out, one for all CBT, one for all PPT, and one for CBT combined with PPT. Table 3.5

contains similar information as in Table 3.3, and Table 3.6 contains similar information as in Table 3.4.

Table 3.5. Raw Score Corresponding to Proficiency Cut for CBT and PPT for Grades 3-12 (Using All Students).

Grade	All (CBT+PPT)	PPT	CBT
3	58	58	57
4	61	61	60
5	63	63	62
6	65	65	64
7	67	67	66
8	69	69	68
9	66	66	66
10	67	67	67
11	68	68	68
12	70	70	70

As shown in Table 3.5, the PPT column is exactly the same as the All column, while for grade 9 to 12, All, PPT, and CBT are all the same. Therefore, only the CBT administrations from grade 3 to grade 8 are different from All, i.e., the classification inconsistency can only happen for CBT students from grade 3 to 8. Therefore, corresponding real data at only those grade levels were checked for the practical impact.

Table 3.6. N Counts of Students Who Met CBT Proficient Cut but not the PPT Proficient Cut (Using All Students).

Grade	N Counts
3	11
4	10
5	13
6	6
7	5
8	9

As shown in Table 3.6, in total, 54 CBT students from grade 3 to grade 8 would be classified differently if calibrated separately vs. calibrated together with the PPT students. Moreover, all such classification inconsistency goes in one direction, i.e., these CBT students are classified as proficient if calibrated separately, but are classified as NOT proficient if calibrated together with the PPT students. An internal discussion was held before the decision was made for reporting purpose. Since the MI-ELPA proficiency cut tend to be on the lenient end, having those 54 students staying in the ELL service (since they are not proficient) a year longer at least would not harm them. Thus, the decision was made to combine data for one single calibration and only produce one conversion table per assessment level.

3.3 Item Level Comparison

In addition to the overall test comparison analyses as described in 3.1 and 3.2 above, some item level analyses were conducted to provide additional information on comparability, again using the matched students. These item level analyses are: (1) Anchor Set and Un-Anchor Set Comparison, (2) Robust Z Statistic of Item Difficulty Estimation, (3) Differential Item Functioning (DIF), and (4) Distribution Homogeneity Test and Item Mean Comparison. All these analyses and corresponding results are described and reported in detail below.

(1) Anchor Set and Un-Anchor Set Comparison

Because the MI-ELPA used the fixed common item parameter approach for equating, the anchor status comparison across the two modes was examined. Since all OP items are the same across two modes, the same anchor set with same anchor values were used for each mode at the initial calibration. After the initial run with WINSTEPS 3.68.2 (Linacre, 2009), displacement values were examined. Each of these values indicates the discrepancy between the anchor values and the estimation of item parameters based on the current data in use. According to Linacre (2009), if a displacement value is greater than or equal to $|0.50|$, the item should be unanchored, i.e., estimated with the current data. After carrying out this rule, we found out that the items need to be excluded from the anchor set are different depending on the mode (see Table 3.7).

Table 3.7. Items Excluded from the Anchor Set for CBT and PPT.

Level	Total Number of Anchor Items	Excluded from CBT	Excluded from PPT
III	26	#52	
IV	28	#22, #41	#7
V	27	#20, #41	

Item 52 at Level III is a 4-point Writing item. Item 7 at Level IV is a Listening item, item 22 is a Reading item, and item 41 is a Writing item. All these three items at Level IV are multiple-choice (MC) items and are all 1-point items. Item 20 at Level V is a Listening item, and item 41 is a Writing item. Both these items at Level V are MC items and are both 1-point items.

In addition to the comparison of items need to be excluded from the anchor set, we also compared item difficulty parameters for all un-anchored items (i.e., items not included in the anchor set). This comparison is done through a scatter plot, with item difficulty parameter estimates from the CBT on the X-axis, and those from the PPT on the Y-axis. A 45-degree line was also plotted. It was expected that all estimations should be clustered relatively tightly around the 45 degree line as well as being randomly distributed around that line.

Figure 3.5.a. Scatter Plot of All Unanchored Items' Item Difficulty Parameter Estimation across Modes—Level III

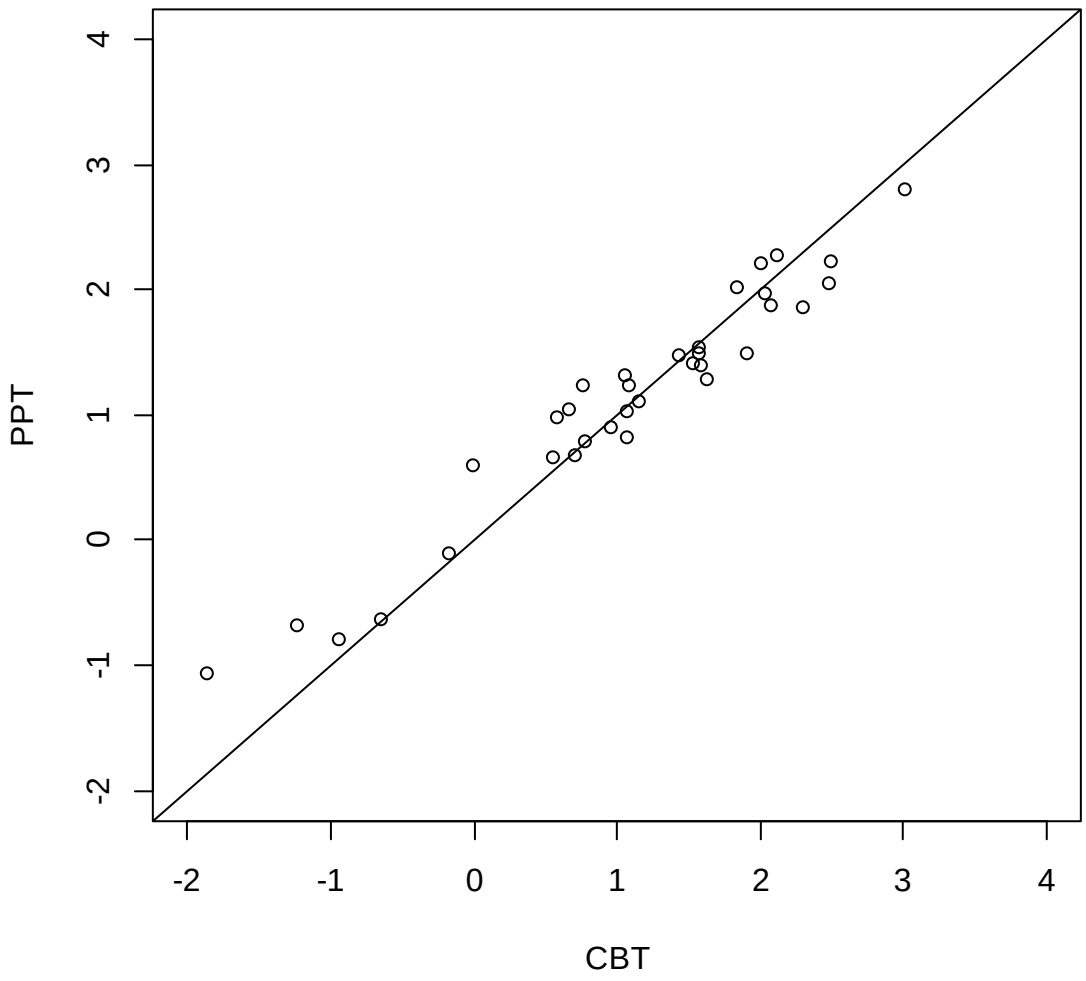


Figure 3.5.b. Scatter Plot of All Unanchored Items' Item Difficulty Parameter Estimation across Modes—Level IV

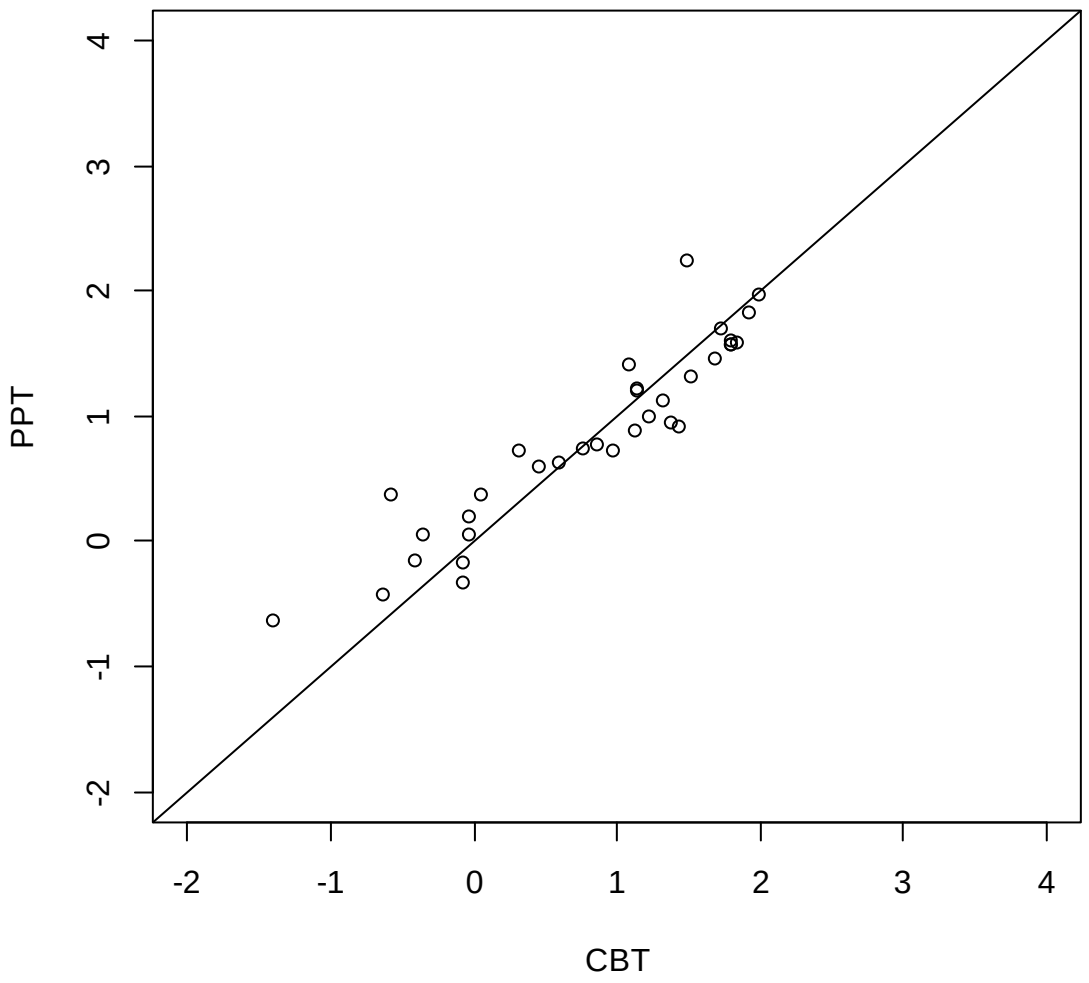
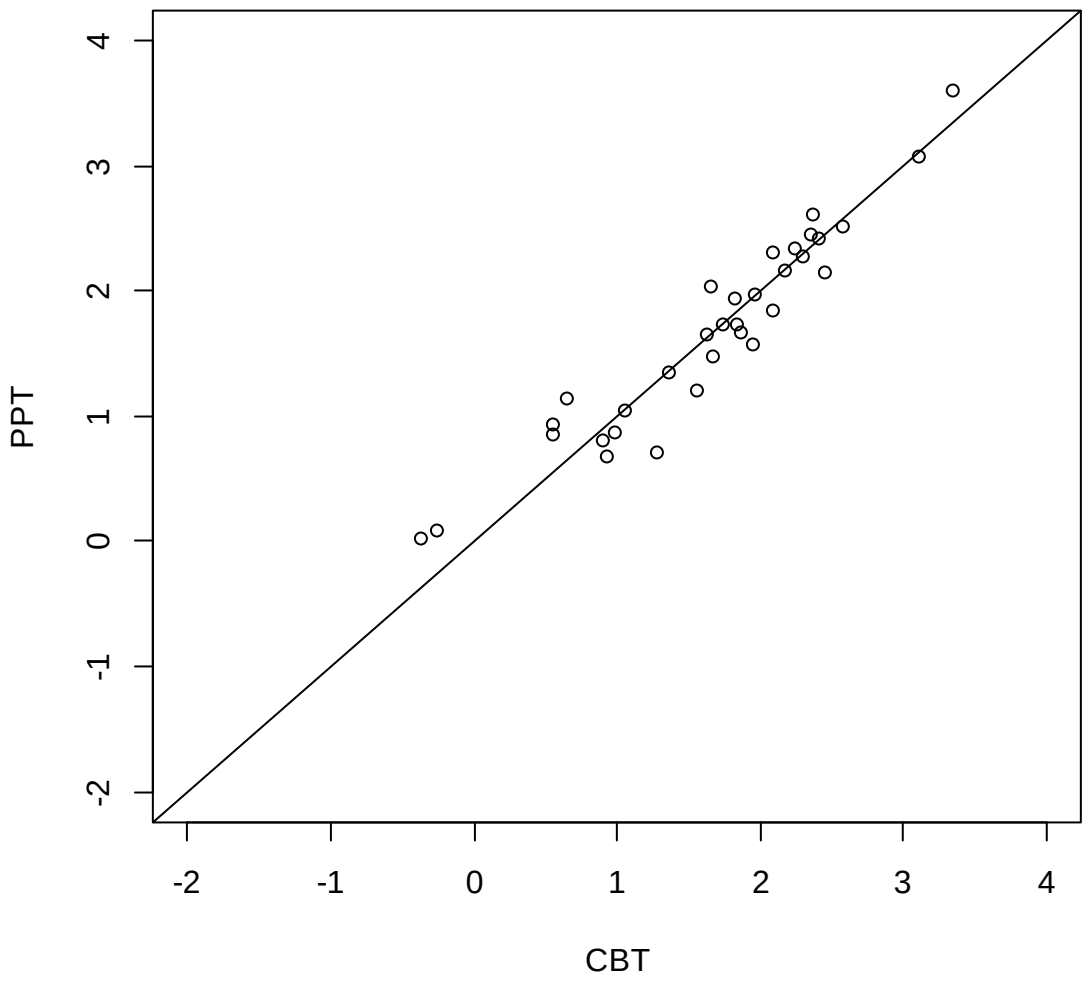


Figure 3.5.c. Scatter Plot of All Unanchored Items' Item Difficulty Parameter Estimation across Modes—Level V



As shown in Figures 3.5.a to 3.5.c, all estimations are clustered relatively tightly around the 45 degree line as well as being randomly distributed around that line. Moreover, the correlation between the two sets of item difficulties (one estimated from the CBT, and the other estimated from the PPT) for each level are: 0.973 for Level III, 0.932 for Level IV, and 0.955 for Level V.

(2) Robust Z Statistic of Item Difficulty Estimation

The formula for the robust Z statistic is as follows:

$$\text{Robust Z} = \frac{D - Median}{0.74(IQR)},$$

where D is the difference between the two sets of item difficulty estimations, and median and IQR are the median and interquartile range of such differences. When absolute value of the robust Z statistic is ≥ 1.645, the differences between item difficulty estimations would be considered significant (Kim & Huynh, 2007).

The robust Z values ranged from -0.196 to 0.366 for Level III, from -0.294 to 0.589 for Level IV, and -0.250 to 0.217 for Level V. None of their absolute values were with the magnitude of ≥1.645, so none of the differences were considered significant.

(3) Differential Item Functioning (DIF)

The DIF analysis for both the dichotomous and polytomous items was conducted with Penfield's (2005) DIFAS software. The details are described below[8].

<u>DIF for Dichotomous Items</u>

For dichotomous items, the Mantel-Haenszel (MH) procedure was used (Mantel & Haenszel, 1959). Specifically, the MH chi-square (χ^2) statistic was used in conjunction with the MH log-odds ratio. These two statistics are discussed in more detail below.

The MH χ^2 indicates "whether there is a relationship between performance on an item and group membership, after taking into account performance on the instrument" (de Ayala, 2009, p. 329). This statistic is calculated as follows (Camilli & Shepard, 1994, p. 120):

$$MH\chi^2 = \frac{\left\{\left|\sum_{j=1}^{S}[A_j - E(A_j)]\right| - 1/2\right\}^2}{\sum_{j=1}^{S} VAR(A_j)} \tag{3.1}$$

where $VAR(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)}$

and $E(A_j) = \frac{n_{Rj} m_{1j}}{T_j}$.

[8] Note that the DIF analyses descriptions are exactly the same as those included in Section 2 in posted technical report.

In Equation (3.1), $A_j - E(A_j)$ represents the discrepancy between the observed number and the expected number of correct responses on the item by the reference group members who have the *j*th score on the matching variable[9]; n_{Rj} and n_{Fj} represent the number of examinees in the reference and focal groups, respectively, for the *j*th score on the matching variable; m_{1j} represents the total number of examinees (both reference and focal) with the *j*th score on the matching variable and with a correct response on the current item; m_{0j} represents the total number of examinees with the *j*th score on the matching variable and with an incorrect response on the current item. The MH χ^2 is evaluated against the standard χ^2 critical values using one degree of freedom.

The MH χ^2 does not indicate the strength of association of the relationship between item performance and group membership. The MH odds ratio can be computed to estimate the strength of this association. The resulting estimate represents the relative likelihood of success on a particular item for members of two different groups of examinees (Camilli, 2006). This odds ratio thus provides an estimate of effect size with a value of 1.0 indicating no DIF. A value greater than 1.0 indicates that, on average, the reference group members performed better than comparable focal group members and a value less than 1.0 indicates that, on average, the reference group members performed worse than comparable focal group members.

The odds of a correct response (proportion passing divided by proportion failing) is *P/Q* (i.e., *P/[1-P])*. The MH odds ratio is simply the odds of a correct response of the reference group divided by the odds of a correct response of the focal group. The formula for its estimation is as follows (Camilli & Shepard, 1994, p. 116):

$$\hat{\alpha}_{MH} = \frac{\sum_{j=1}^{S} A_j D_j / T_j}{\sum_{j=1}^{S} B_j C_j / T_j} \tag{3.2}$$

where $S = K-1$ and represents the actual number of 2 x 2 contingency tables (assuming the tables have at least 1 person in each cell); *K* represents the number of items on the test; *j* signifies the *j*th score on the matching variable and runs from 0 to *K*[10]. For *j*th score category, A_j represents the number of reference group members with a correct response, B_j represents the number of reference group members with an incorrect response, C_j represents the number of focal group members with a correct response, and D_j represents the number of focal group members with an incorrect response. T_j represents the total number of examinees who have the *j*th score on the matching variable.

The corresponding null hypothesis is that the odds of getting the item correct are equal for the two groups (i.e., the odds ratio is equal to 1):

$$H_0\text{: } \alpha_{MH} = 1 \tag{3.3}$$

---

[9] Total observed score is used as the matching variable for DIF analysis here.

[10] Although the value of the matching variable runs from 0 to *K*, the all correct (*K*) and all incorrect (0) score categories are not included in the DIF analysis in order to avoid having a denominator equal to 0.

In order to make the odds ratio symmetrical around zero with its range located in the interval $-\infty$ to $+\infty$, the odds ratio is transformed into a log-odds ratio as follows (Camilli & Shepard, 1994, p.116):

$$\hat{\lambda}_{MH} = \log(\alpha_{MH}) \tag{3.4}$$

The natural logarithm transformation of this odds ratio is symmetrical around zero (where 0 indicates no DIF). This DIF measure is a signed index where a positive value represents DIF in favor of the reference group and a negative value indicates DIF in favor of the focal group.

The variance of the log-odds ratio estimate (V_λ) is computed as follows (Camilli & Shepard, 1994, p. 121):

$$V_\lambda = \frac{\sum_{j=1}^{S} T_j^{-2}(A_j D_j + \alpha_{MH} B_j C_j)[A_j + D_j + \alpha_{MH}(B_j + C_j)]}{2(\sum_{j=1}^{S} A_j D_j / T_j)^2}. \tag{3.5}$$

The terms included in Equation (3.5) correspond to those presented for Equation (3.2). In practice, a standardized MH log-odds ratio is computed by dividing the estimate $\hat{\lambda}_{MH}$ by the estimated standard error. According to Penfield (2007, p.16), "A value greater than 2.0 or less than -2.0 may be considered evidence of the presence of DIF."

Table 3.8 depicts the DIF classifications for theMI-ELPA MC items[11].

Table 3.8 DIF Classification for MC Items

Category	Description	Criterion
A	Negligible DIF	If either MH common log-odds ratio ($\hat{\lambda}_{MH}$) is not significantly different from zero or $\lvert\hat{\lambda}_{MH}\rvert < 0.426$
B	Moderate DIF	If $\hat{\lambda}_{MH}$ is significantly different from zero and $\lvert\hat{\lambda}_{MH}\rvert \geq 0.426$, and either (a) $\lvert\hat{\lambda}_{MH}\rvert \leq 0.638$, or (b) $\lvert\hat{\lambda}_{MH}\rvert$ is not significantly greater than 0.426
C	Large DIF	If $\lvert\hat{\lambda}_{MH}\rvert$ is significantly greater than 0.426 and $\lvert\hat{\lambda}_{MH}\rvert > 0.638$

Note. Adapted from Penfield (2007).

[11] Note that these rules are implemented at the back end of the DIFAS program for computing the "ETS" categorization values.

DIF for Polytomous Items

Since the MI-ELPA includes CR items that are polytomously scored, the MH procedure cannot be used for these items. Instead, an extension of the MH χ^2 procedure was computed (see Mantel, 1963) and a generalization of the MH odds ratio for ordered categories, the Liu-Agresti Cumulative Common Odds Ratio (Liu & Agresti, 1996), was used to assess DIF for the polytomous items.

Mantel (1963) proposed an extension of the MH procedure for ordered response categories that involved comparing the performance of the reference and focal groups, conditional on a matching variable. The statistic is computed as follows (Zwick, Donaghue, & Grima, 1993, p. 239):

$$Mantel\ \chi^2 = \frac{(\sum_k F_k - \sum_k E(F_k))^2}{\sum_k VAR(F_k)} \tag{3.6}$$

where F_k is the sum of scores for the focal group at the *k*th level of the matching variable and is defined as

$$F_k = \sum_t y_t n_{Ftk},$$

and the expectation of F_k under the hypothesis of no association is

$$E(F_k) = \frac{n_{F+k}}{n_{++k}} \sum_t y_t\, n_{+tk}\ ,$$

and the variance of F_k under the assumption of no association is

$$\mathrm{Var}(\mathrm{F_k}) = \frac{\mathrm{n_{R+k} n_{F+k}}}{\mathrm{n_{++k}^2(n_{++k} - 1)}} \left\{ \left(\mathrm{n_{++k}} \sum_{\mathrm{t}} \mathrm{y_t^2 n_{+tk}} \right) - \left(\sum_{\mathrm{t}} \mathrm{y_t\, n_{+tk}} \right)^2 \right\}. \tag{3.7}$$

Using the Mantel approach for ordered categories, the data are organized into a 2 x *T* x *K* contingency table, where *T* is the number of response categories and *K* is the number of levels of the matching variable. $y_1, y_2, \ldots, y_T$ represent the *T* scores that can be obtained on the item; n_{Rtk} and n_{Ftk} represent the number of examinees in the reference and focal groups, respectively, who are at the *k*th level of the matching variable and received an item score of y_t on the item. The "+" denotes summation over a particular index (e.g., n_{R+k} denotes the total number of reference group members at the *k*th level of the matching variable). Under the null hypothesis of no association, the Mantel statistic has a chi-square distribution with one degree of freedom. For dichotomous items, the Mantel statistic reduces to the MH statistic (without the continuity correction).

The Liu-Agresti estimator of the common odds ratio has the following mathematical formulation (Liu & Agresti, 1996; Penfield & Algina, 2003, p. 357):

$$\hat{\psi}_{LA} = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J-1} A_{jk}D_{jk}/N_k}{\sum_{k=1}^{K}\sum_{j=1}^{J-1} B_{jk}C_{jk}/N_k}, \tag{3.8}$$

where each item has J levels of response scores ($j = 1, \ldots, J$, and thus $(J-1)$ dichotomizations), the matching variable has K levels ($k = 1, \ldots, K$), and N_k is the total number of individuals at level k. The cumulative frequency for response level j at stratum k for the reference group is denoted as $N^*_{Rjk} = N_{R1k} + \cdots + N_{Rjk}$ and that of the focal group is denoted as $N^*_{Fjk} = N_{F1k} + \ldots + N_{Fjk}$. Moreover, $A_{jk} = N^*_{Rjk}, B_{jk} = N_{Rk} - N^*_{Rjk}, C_{jk} = N^*_{Fjk}$, and $D_{jk} = N_{Fk} - N^*_{Fjk}$. These values are analogous to A_j, B_j, C_j and D_j in Equation (3.2), but it is important to note that in the dichotomous case A and C are associated with the correct responses but for the cumulative polytomous case A and C are associated with the lower portion of the response variable dichotomization. Thus, for dichotomous items, the Liu-Agresti statistic reduces to the inverse of $\hat{\alpha}_{MH}$. To match the scale of the Liu-Agresti estimator to the scale for dichotomous items, the following computation is necessary (Penfield & Algina, 2003, p. 357):

$$\hat{\alpha}_{LA} = \frac{1}{\hat{\psi}_{LA}} \tag{3.9}$$

Values equal to 1.0 indicate no DIF while values less than 1.0 indicate DIF in favor of the focal group. Similar to the MH procedure, the natural log of $\hat{\alpha}_{LA}$ can be computed, where $\log(\hat{\alpha}_{LA})$ of 0 indicates no DIF, $\log(\hat{\alpha}_{LA})$ greater than 0 indicates DIF against the focal group, and $\log(\hat{\alpha}_{LA})$ less than 0 indicates DIF against the reference group. In practice, a standardized Liu-Agresti cumulative common log-odds ratio is computed by dividing $\log(\hat{\alpha}_{LA})$ by the estimated standard error. According to Penfield (2007, p.18), "A value greater than 2.0 or less than -2.0 may be considered evidence of the presence of DIF."

The Liu-Agresti estimate of the variance of $\log(\hat{\psi}_{LA})$ is expressed as follows (Penfield & Algina, 2003, p. 357):

$$\mathrm{Var}\left[\log\left(\hat{\psi}_{LA}\right)\right] = \frac{\sum_{k=1}^{K}\hat{\xi}_k(\hat{\psi})}{\hat{\psi}_{LA}^2\left[\sum_{k=1}^{K}\sum_{j=1}^{J-1} B_{jk}C_{jk}/N_k\right]^2}, \tag{3.10}$$

where $$\hat{\xi}_k(\hat{\psi}) = \sum_{j=j'=1}^{J-1}\hat{\phi}_{jj'k}(\hat{\psi}) + 2\sum_{j<j'}^{J-1}\hat{\phi}_{jj'k}(\hat{\psi}),$$

and $$\hat{\phi}_{jj'k}(\hat{\psi}) = \frac{n_{Rk}n_{Fk}}{N_k^2}\left\{\frac{\hat{\psi}B_{jk}C_{j'k}}{n_{Rk}}\left[1 + (\hat{\psi}-1)\frac{C_{j'k}}{n_{Rk}}\right] + \frac{A_{jk}D_{j'k}}{n_{Fk}}\left[\hat{\psi} - (\hat{\psi}-1)\frac{A_{jk}}{n_{Rk}}\right]\right\}.$$

$$j \le j' = 1, \ldots, J-1$$

This variance is a generalization of the variance estimator for the MH estimate of the common log-odds ratio for dichotomous items. Because $\log(\hat{\alpha}_{LA}) = -\log(\hat{\psi}_{LA})$, the statistics have the same asymptotic variance (Penfield & Algina, 2003).

Because the Liu-Agresti estimator is theoretically equivalent to the MH odds ratio, a classification scheme can be constructed that parallels the ETS scheme for dichotomous items (Penfield, 2007). Table 3.9 presents the DIF classification for the MI-ELPA polytomous CR items[12]. Note that the rules used in Table 3.9 are identical to those used in Table 3.8, with the only exception being the statistics the rules are applied to.

Table 3.9 DIF Classification for Polytomous CR Items

Category	Description	Criterion
AA	Negligible DIF	If either the Liu-Agresti cumulative common log-odds ratio ($\log(\hat{\alpha}_{LA})$ is not significantly different from zero or $\lvert\log(\hat{\alpha}_{LA})\rvert < 0.426$
BB	Moderate DIF	If $\log(\hat{\alpha}_{LA})$ is significantly different from zero and $\lvert\log(\hat{\alpha}_{LA})\rvert \geq 0.426$, and either (a) $\lvert\log(\hat{\alpha}_{LA})\rvert \leq 0.638$ or (b)$\lvert\log(\hat{\alpha}_{LA})\rvert$ is not significantly greater than 0.426
CC	Large DIF	If $\lvert\log(\hat{\alpha}_{LA})\rvert$ is significantly greater than 0.426 and $\lvert\log(\hat{\alpha}_{LA})\rvert > 0.638$

Note. Adapted from Penfield (2007).

The results of DIF analysis per assessment level are reported in Table 3.10. All three B or BB items at Level III favored the CBT students (i.e., the CBT students performed better on those items than the PPT students), while the one C or CC item at Level III favored PPT students. This one item is a 4-point Writing item, so it may have something to do with the typing skills for students at grades 3 to 5. At Level IV, again, the two B or BB items favored the CBT students, and the one C or CC item (again a 4-point Writing item) favored the PPT students. At Level V, two B or BB items (both are MC items) favored the CBT students, while the rest B or BB items and the one C or CC item all favored the PPT students and are all Speaking items. Speaking is a one-on-one test, so this may or may not have something to do with teachers' familiarity with the two computers set up.

Table 3.10. DIF Analysis Results

Level	Total # of Items	# of B or BB Items	# of C or CC Items
III	59	3	1
IV	59	2	1
V	60	4	1

[12] There are 1-point CR items at Level I. In order to identify these as CR items, AA, BB, and CC are assigned and the analysis is conducted as if they were MC items (i.e., dichotomous items). For these items, the rules in Table 3.8 were applied instead.

(4) Distribution Homogeneity Test and Item Mean Comparison

In order to make the distribution homogeneity test meaningful, the omit and multiple marks for each MC item were excluded before the homogeneity test was carried out for each MC item, and only scores from CR items were considered for the homogeneity test (i.e., different rater codes were ignored, as those responses are all scored as 0 in the end). T-test on item means was carried out for each item across the two modes. The results are reported in Tables 3.11.

Table 3.11. Number of Items Showing Significant Differences (at α=0.05) across Modes

Level	Total # of Items	# of Items Showing Significant Differences	
		Homogeneity Test	T-Test
III	59	18	13
IV	59	4	5
V	60	11	8

According to Table 3.11, different numbers of items were shown to be statistically significant depending on what statistical test was used. Moreover, only a small portion of items were consistently identified by the two tests. Specifically, 11 items were commonly identified at Level III, 2 items were commonly identified at Level IV, and only 4 items were commonly identified at Level V.

3.4 Brief Summary

In conclusion, at the overall test level, especially when the focus is on classification consistency and when practical impact was taken into consideration, we considered all three assessments being comparable across the two delivery modes. However, overall comparability does not equivalent to item level comparability, as shown by the above four sets of item level analyses. In addition, different analyses identify different problematic items. The only item identified as different across modes by majority of the analyses (except Robust Z statistic, as this one identifies no problematic items at all) is #52 at Level III. In addition to this item, the items identified by DIF analyses, homogeneity test, and the t-test simultaneously are: #12 and #48 at Level III, #2 and #52 at Level IV, and #9 and #54 at Level V. These items, since being identified by multiple statistical tests, seem to worth a closer examination of item appearance on the computer screen and on paper[13].

[13] Since all these items' computer screen appearances were not documented, such information was lost. Therefore, such content related comparison is no longer possible.

SECTION 4. CONCLUSIONS AND DISCUSSION

This section summarizes briefly the procedures and findings in this report and further discusses the issue of possible violation of ignorable treatment assignment assumption.

4.1 Conclusions

In this report, we tried to include all previous achievement scores (including the scores on the previous year's MI-ELPA) in the propensity score estimation model. 2011 was the first year the CBT was available, and because of the need for administering the Speaking domain on a one-on-one basis (thus two computers are required during testing for each individual student), 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 to be the ones to participate. We also included student demographic variables as the ELL population is very diverse, and we hoped that by including those variables as covariates, it would help to remove some bias with regard to different computer-person interaction due to culture differences.

The majority of the analyses results reported in Section 3 are based on the matched sample obtained using estimated propensity scores. As mentioned at the end of Section 3, even though the overall test could be considered as comparable especially from practical perspective, item level analyses indicated that some items still demonstrated mode differences. Closer examination of such items delivered by different modes are needed, and if no obvious reasons can be found to explain such differences, how to handle these items in the future may need further investigation.

Even though we have included most of the variables researchers have identified as useful in this scenario, we did not manage to include one variable identified by researchers such as Bennett, et al. (2008) and Lottridge, et al. (2011): student computer skill scores. This information has never been collected by the Michigan Department of Education at the state level so far. Whether this will pose a validity threat on the conclusions is discussed below.

4.2 Discussion

In this sub-section, we would like to argue that not including student computer skill scores would not alter our conclusions found in Section 3 that the overall tests are comparable across modes.

Because computers are becoming more widely used in personal life around the world, we did not expect such skills to be too different among ELLs coming from different countries, especially because students took the practice test and were allowed to back out of taking the CBT and take the PPT instead if they felt uncomfortable with it. 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 this variable 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. Since the PPT is the most familiar test delivery mode to students around the world, taking the CBT could only possibly harm student scores rather than boosting student performance. Also, students willing to take the CBT felt comfortable with using computers, and thus must have had experience in using them and should have had opportunities to use them. Such students were more likely to have higher SES. By the inclusion of the variable "economically disadvantaged," we have taken computer skills into consideration indirectly. Therefore, the missing of the computer skill variable in the propensity score estimation should not alter the reported mode comparison results.

In conclusion, the decision made at the scoring stage that the CBT student can be treated as if they took the PPT and the item calibration and scaling procedures being carried out using combined data was appropriate.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *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 March 30, 2013 from http://www.jtla.org

Bowers, J., Fredrickson, M., & Hansen, B. (2010). RItools: Randomization Inference Tools. R package version 0.1-11.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*, 1149-1156.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31-72.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 221-256). Westport, CT: American Council on Education and Praeger.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: SAGE Publications.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

de Ayala, R. J. (2009). *The theory and practice of Item Response Theory*. New York: Guilford Press.

Hansen, B. B. & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science, 23*(2):219--236.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association, 99(467), 609-618.*

Hansen, B.B. and 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*(3), 234-249.

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*(475), 901-910.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453.

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 March 30, 2013 from http://www.jtla.org

Linacre, J. M. (2009). *WINSTEPS* (Version 3.68.2) [Computer Software]. Chicago: Winsteps.com.

Liu, I-M., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*, 1223-1234.

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.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *American Statistical Association Journal, 58*, 690-700.

Michigan Department of Education (2011). Michigan English language proficiency assessment: Final technical report 2011 administration. Retrieved March 30, 2013 from http://www.michigan.gov/documents/mde/Final_Acrobat7_317649_7.pdf

Muthén, L. K., & Muthén, B. O. (2009). *Mplus* (Version 5.21) [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*(1), 71-83.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement, 29,* 150-151.

Penfield, R. D. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education, 20*, 335-355.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*, 353-370.

Pomplun, M., Ritchie, T., & Custer, M. (2006). Factors in paper-and-pencil and computer reading score differences at the primary grades. *Educational Assessment, 11*(2), 127-143.

Rosenbaum, P. R. (2010). *Observational studies* (2nd Ed.). New York: Springer-Verlag.

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*(5), 849-869.

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*(3), 250-267.

Stuart, E. A. (2010). Matching methods for causal inference: *A review and a look forward. Statistical Science, 25*(1), 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., Lin, C.-H., & Kong, J. (March, 2008). *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.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Westport, Ct: American Council on Education and Praeger.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233-251.

APPENDIX A: DIFFERENCES IN COVARIATES' MEANS AND PROPORTIONS BEFORE AND AFTER MATCHING

A.1: Grade 3

	Before Matching		After Matching	
	CBT (N=308)	PPT (N=6,603)	CBT (N=285)	PPT (N=285)
Individual Level Variables				
Native Language: Spanish	9.09%	36.15%***	9.82%	7.37%
Native Language: Arabic	8.44%	26.09%***	9.12%	7.37%
Economically Disadvantaged	76.30%	77.65%	75.44%	78.25%
Disabled	8.44%	10.03%	8.77%	11.93%
Female	47.08%	46.31%	45.26%	45.21%
Ethnicity: Black	3.90%	2.70%	4.21%	2.81%
Ethnicity: Hispanic	29.55%	43.30%***	27.72%	27.37%
Ethnicity: White	45.78%	39.39%*	46.67%	45.96%
Ethnicity: Asian	19.81%	13.90%**	20.35%	21.75%
ELPA Proficient in 2010	26.30%	29.08%	26.67%	25.96%
Math Proficient in 2010	14.29%	19.10%*	13.68%	14.74%
Reading Proficient in 2010	29.22%	31.73%	28.07%	29.47%
ELPA Scale Score in 2010	578.8	579.9 [a]	578.7	579.1
Math Scale Score in 2010	319.3	320.8 [a]	319.0	318.9
Reading Scale Score in 2010	311.3	312.7 [a]	311.1	311.2

	Before Matching		After Matching	
	CBT (N=308)	PPT (N=6,603)	CBT (N=308)	PPT (N=6,603)
School Level Variables				
Number of ELL	15.3	35.8*** [a]	15.2	15.0 [a]
Number of Computers with High Speed Internet per ELL	15.2	14.6	15.5	14.1
Average of Mean ELPA Scale Score in Last Three Years	606.0	606.0	605.5	604.0
Average of Mean Math Scale Score in Last Three Years	320.2	322.6*** [a]	320.2	320.0
Average of Mean Reading Scale Score in Last Three Years	314.4	316.3*** [a]	314.2	315.1
Average of Proportion of Proficient on ELPA in Last Three Years	30.86	29.84 [a]	30.09	28.40
Average of Proportion of Proficient on Math in Last Three Years	18.01	22.30*** [a]	17.65	17.53
Average of Proportion of Proficient on Reading in Last Three Years	33.23	35.80** [a]	32.62	34.06

Note. CBT=computer-based test. PPT=paper-based test. * $p<.05$, ** $p<.01$, *** $p<.001$.
[a] Variance significantly different at $p<.05$.

A.2: Grade 4

	Before Matching		After Matching	
	CBT	PPT	CBT	PPT
	(N=243)	(N=5,774)	(N=190)	(N=190)
Individual Level Variables				
Native Language: Spanish	9.88%	39.33%***	1.05%	3.68%
Native Language: Arabic	5.35%	23.07%***	6.84%	3.16%
Economically Disadvantaged	73.66%	78.30%	69.47%	68.42%
Disabled	10.29%	12.35%	11.58%	15.26%
Female	39.51%	46.81%*	39.47%	36.32%
Ethnicity: Black	2.88%	2.34%	3.16%	2.63%
Ethnicity: Hispanic	30.04%	46.48%***	18.42%	20.53%
Ethnicity: White	43.21%	36.68%*	53.16%	50.53%
Ethnicity: Asian	22.63%	13.82%***	23.68%	25.79%
ELPA Proficient in 2010	20.99%	19.85%	17.37%	16.84%
Math Proficient in 2010	18.93%	18.88%	16.84%	20.00%
Reading Proficient in 2010	27.98%	31.99%	27.89%	24.74%
Writing Proficient in 2010	18.93%	25.44%*	21.05%	17.37%
ELPA Scale Score in 2010	600.8	599.8	598.5	598.0
Math Scale Score in 2010	419.7	418.1	418.4	419.5
Reading Scale Score in 2010	409.5	409.9	408.5	408.9
Writing Scale Score in 2010	382.9	384.8 [a]	382.4	382.1

	Before Matching		After Matching	
	CBT (N=243)	PPT (N=5,774)	CBT (N=243)	PPT (N=5,774)
School Level Variables				
Number of ELL	13.9	29.4*** [a]	12.7	13.0
Number of Computers with High Speed Internet per ELL	20.3	16.5 [a]	23.5	22.6 [a]
Average of Mean ELPA Scale Score in Last Three Years	622.2	619.7** [a]	622.2	620.6
Average of Mean Math Scale Score in Last Three Years	418.5	418.4 [a]	417.7	418.4
Average of Mean Reading Scale Score in Last Three Years	408.9	409.3 [a]	408.0	408.3
Average of Mean Writing Scale Score in Last Three Years	383.3	386.0***	382.9	384.1
Average of Proportion of Proficient on ELPA in Last Three Years	40.31	43.86*	42.90	40.05
Average of Proportion of Proficient on Math in Last Three Years	20.89	20.03	19.64	20.82
Average of Proportion of Proficient on Reading in Last Three Years	28.87	31.75*	28.73	27.45
Average of Proportion of Proficient on Writing in Last Three Years	19.95	25.88***	20.06	20.17

Note. CBT=computer-based test. PPT=paper-based test. * $p<.05$, ** $p<.01$, *** $p<.001$.
[a] Variance significantly different at $p<.05$.

A.3: Grade 5

	Before Matching		After Matching	
	CBT	PPT	CBT	PPT
	(N=271)	(N=4,772)	(N=211)	(N=211)
Individual Level Variables				
Native Language: Spanish	8.12%	40.09%***	1.42%	3.79%
Native Language: Arabic	4.06%	22.17%***	5.21%	5.21%
Economically Disadvantaged	69.00%	79.67%***	64.93%	68.25%
Disabled	12.92%	13.33%	14.69%	16.11%
Female	47.23%	47.53%	44.55%	44.08%
Ethnicity: Black	2.95%	2.77%	3.79%	2.84%
Ethnicity: Hispanic	26.57%	47.88%***	18.96%	22.27%
Ethnicity: White	44.28%	35.35%**	49.29%	49.76%
Ethnicity: Asian	25.83%	12.78%***	27.49%	25.12%
ELPA Proficient in 2010	28.41%	31.33%	28.44%	31.28%
Math Proficient in 2010	23.99%	17.64%*	23.70%	27.49%
Reading Proficient in 2010	25.09%	28.79%	26.07%	27.96%
Science Proficient in 2010	1.48%	1.91%	1.90%	1.42%
ELPA Scale Score in 2010	611.8	611.6 [a]	610.6	614.1
Math Scale Score in 2010	516.4	510.8** [a]	516.0	516.7
Reading Scale Score in 2010	507.9	506.9 [a]	506.8	509.6
Science Scale Score in 2010	502.4	499.9	502.1	503.2

	Before Matching		After Matching	
	CBT (N=271)	PPT (N=4,772)	CBT (N=271)	PPT (N=4,772)
School Level Variables				
Number of ELL	16.3	27.7*** [a]	15.6	17.1 [a]
Number of Computers with High Speed Internet per ELL	19.1	17.8	21.5	18.4 [a]
Average of Mean ELPA Scale Score in Last Three Years	631.0	628.7** [a]	630.1	630.5
Average of Mean Math Scale Score in Last Three Years	518.0	510.7*** [a]	517.5	516.8
Average of Mean Reading Scale Score in Last Three Years	507.6	506.5 [a]	507.5	508.3
Average of Mean Science Scale Score in Last Three Years	503.8	502.3** [a]	504.2	503.9
Average of Proportion of Proficient on ELPA in Last Three Years	53.10	47.99*** [a]	51.66	51.90
Average of Proportion of Proficient on Math in Last Three Years	27.79	18.91*** [a]	27.64	26.74
Average of Proportion of Proficient on Reading in Last Three Years	26.75	26.72	27.38	27.91
Average of Proportion of Proficient on Science in Last Three Years	2.14	2.14 [a]	2.19	2.13 [a]

Note. CBT=computer-based test. PPT=paper-based test. * $p<.05$, ** $p<.01$, *** $p<.001$.
[a] Variance significantly different at $p<.05$.

A.4: Grade 6

	Before Matching		After Matching	
	CBT	PPT	CBT	PPT
	(N=144)	(N=3,902)	(N=112)	(N=112)
Individual Level Variables				
Native Language: Spanish	13.89%	35.06%***	8.93%	7.14%
Native Language: Arabic	4.17%	24.01%***	5.36%	3.57%
Economically Disadvantaged	66.67%	83.09%***	66.07%	66.96%
Disabled	7.64%	14.40%*	8.93%	8.93%
Female	52.78%	44.72%	47.32%	47.32%
Ethnicity: Black	5.56%	3.02%	7.14%	2.68%
Ethnicity: Hispanic	49.31%	44.85%	43.75%	44.64%
Ethnicity: White	16.67%	38.95%***	16.96%	15.18%
Ethnicity: Asian	27.08%	12.35%***	30.36%	35.71%
ELPA Proficient in 2010	42.36%	40.13%	40.18%	36.61%
Math Proficient in 2010	22.22%	14.22%*	20.54%	22.32%
Reading Proficient in 2010	23.61%	19.73%	23.21%	21.43%
Social Studies in 2010	8.33%	4.95%	8.04%	8.04%
ELPA Scale Score in 2010	624.2	621.0 [a]	621.6	619.1
Math Scale Score in 2010	617.0	609.8*** [a]	615.8	615.9
Reading Scale Score in 2010	601.7	600.2	600.7	600.9
Social Studies Scale Score in 2010	600.8	597.6*	599.9	600.5

	Before Matching		After Matching	
	CBT (N=144)	PPT (N=3,902)	CBT (N=144)	PPT (N=3,902)
School Level Variables				
Number of ELL	10.5	39.7*** [a]	10.7	10.5
Number of Computers with High Speed Internet per ELL	37.9	19.9*** [a]	39.8	48.2 [a]
Average of Mean ELPA Scale Score in Last Three Years	618.0	613.4***	616.5	616.8
Average of Mean Math Scale Score in Last Three Years	619.6	610.7*** [a]	618.0	615.9
Average of Mean Reading Scale Score in Last Three Years	606.5	604.1**	605.4	605.4
Average of Mean Social Studies Scale Score in Last Three Years	598.1	596.7*	597.8	597.3
Average of Proportion of Proficient on ELPA in Last Three Years	25.46	20.38*** [a]	23.05	25.38
Average of Proportion of Proficient on Math in Last Three Years	28.91	17.66*** [a]	26.90	25.00
Average of Proportion of Proficient on Reading in Last Three Years	30.48	24.94***	29.59	30.67
Average of Proportion of Proficient on Social Studies in Last Three Years	8.13	6.85	7.96	8.89

Note. CBT=computer-based test. PPT=paper-based test. * $p<.05$, ** $p<.01$, *** $p<.001$.
[a] Variance significantly different at $p<.05$.

A.5: Grade 7

	Before Matching		After Matching	
	CBT	PPT	CBT	PPT
	(N=121)	(N=3,721)	(N=100)	(N=100)
Individual Level Variables				
Native Language: Spanish	18.18%	35.31%***	20.00%	25.00%
Native Language: Arabic	2.48%	24.81%***	3.00%	4.00%
Economically Disadvantaged	61.98%	81.40%***	61.00%	65.00%
Disabled	11.57%	13.28%	11.00%	8.00%
Female	44.63%	46.73%	43.00%	47.00%
Ethnicity: Black	1.65%	3.14%	2.00%	0.00%
Ethnicity: Hispanic	52.07%	42.62%*	51.00%	56.00%
Ethnicity: White	8.26%	40.20%***	10.00%	7.00%
Ethnicity: Asian	37.19%	13.22%***	36.00%	36.00%
ELPA Proficient in 2010	21.49%	15.88%	19.00%	17.00%
Math Proficient in 2010	27.27%	13.44%***	24.00%	20.00%
Reading Proficient in 2010	18.18%	15.96%	16.00%	16.00%
Writing Proficient in 2010	19.83%	16.61%	19.00%	13.00%
ELPA Scale Score in 2010	612.2	608.5	611.5	610.3
Math Scale Score in 2010	720.0	708.7*** [a]	717.6	716.1
Reading Scale Score in 2010	700.4	696.4	700.2	699.5
Writing Scale Score in 2010	678.4	678.3	678.3	675.0

	Before Matching		After Matching	
	CBT (N=121)	PPT (N=3,721)	CBT (N=121)	PPT (N=3,721)
School Level Variables				
Number of ELL	15.79	37.74*** [a]	15.01	13.66
Number of Computers with High Speed Internet per ELL	40.60	20.55*** [a]	42.60	36.21 [a]
Average of Mean ELPA Scale Score in Last Three Years	622.8	621.5 [a]	622.6	622.5
Average of Mean Math Scale Score in Last Three Years	721.6	709.9*** [a]	719.5	719.5
Average of Mean Reading Scale Score in Last Three Years	704.8	700.8*** [a]	704.2	705.5 [a]
Average of Mean Writing Scale Score in Last Three Years	680.1	680.6	679.4	679.7
Average of Proportion of Proficient on ELPA in Last Three Years	28.36	26.65 [a]	28.51	28.42
Average of Proportion of Proficient on Math in Last Three Years	32.72	15.46*** [a]	28.80	27.59
Average of Proportion of Proficient on Reading in Last Three Years	27.73	19.75*** [a]	25.89	26.95 [a]
Average of Proportion of Proficient on Writing in Last Three Years	25.34	17.61** [a]	21.26	19.59

Note. CBT=computer-based test. PPT=paper-based test. * $p<.05$, ** $p<.01$, *** $p<.001$.
[a] Variance significantly different at $p<.05$.

A.6: Grade 8

	Before Matching		After Matching	
	CBT	PPT	CBT	PPT
	(N=136)	(N=3,559)	(N=109)	(N=109)
Individual Level Variables				
Native Language: Spanish	23.53%	37.62%***	21.10%	21.10%
Native Language: Arabic	8.09%	24.16%***	10.09%	8.26%
Economically Disadvantaged	60.29%	79.99%***	64.22%	59.63%
Disabled	11.03%	12.39%	11.93%	9.17%
Female	43.38%	46.05%	44.04%	48.62%
Ethnicity: Black	2.21%	3.12%	1.83%	0.00%
Ethnicity: Hispanic	47.06%	45.77%	44.04%	44.04%
Ethnicity: White	25.74%	38.41%**	28.44%	27.52%
Ethnicity: Asian	25.00%	12.11%***	25.69%	28.44%
ELPA Proficient in 2010	25.74%	20.74%	23.85%	29.36%
Math Proficient in 2010	19.12%	8.91%***	17.43%	23.85%
Reading Proficient in 2010	18.38%	18.46%	20.18%	18.35%
Science Proficient in 2010	0.74%	1.52%	0.92%	1.83%
ELPA Scale Score in 2010	620.6	616.7 [a]	620.2	623.8
Math Scale Score in 2010	810.6	803.0*** [a]	809.7	812.0
Reading Scale Score in 2010	802.6	801.5 [a]	802.4	803.3
Science Scale Score in 2010	801.0	801.0	801.3	801.3

	Before Matching		After Matching	
	CBT (N=136)	PPT (N=3,559)	CBT (N=136)	PPT (N=3,559)
School Level Variables				
Number of ELL	12.00	40.78*** [a]	11.1	11.8 [a]
Number of Computers with High Speed Internet per ELL	40.10	21.32*** [a]	41.9	32.9 [a]
Average of Mean ELPA Scale Score in Last Three Years	632.0	628.8*** [a]	631.6	632.1
Average of Mean Math Scale Score in Last Three Years	812.9	804.1*** [a]	812.4	814.1
Average of Mean Reading Scale Score in Last Three Years	802.7	802.3 [a]	802.9	804.0 [a]
Average of Mean Science Scale Score in Last Three Years	803.2	800.2*** [a]	803.3	803.4
Average of Proportion of Proficient on ELPA in Last Three Years	29.63	27.19 [a]	28.66	29.18
Average of Proportion of Proficient on Math in Last Three Years	21.95	10.96*** [a]	21.68	24.27
Average of Proportion of Proficient on Reading in Last Three Years	20.76	20.72	20.94	21.81
Average of Proportion of Proficient on Science in Last Three Years	2.11	2.06 [a]	2.06	2.25

Note. CBT=computer-based test. PPT=paper-based test. * $p<.05$, ** $p<.01$, *** $p<.001$.
[a] Variance significantly different at $p<.05$.

A.7: Grade 9

	Before Matching		After Matching	
	CBT	PPT	CBT	PPT
	(N=196)	(N=4,010)	(N=4,010)	(N=196)
Individual Level Variables				
Native Language: Spanish	13.78%	34.89%***	15.79%	18.13%
Native Language: Arabic	5.10%	28.98%***	5.85%	5.26%
Economically Disadvantaged	76.02%	78.65%	74.27%	73.10%
Disabled	4.59%	11.02%**	5.26%	5.26%
Female	48.47%	43.69%	48.54%	51.46%
Ethnicity: Black	5.10%	4.06%	5.85%	4.68%
Ethnicity: Hispanic	41.33%	40.40%	36.26%	33.33%
Ethnicity: White	32.65%	41.32%*	36.26%	40.94%
Ethnicity: Asian	20.92%	13.52%**	21.64%	21.05%
ELPA Proficient in 2010	11.73%	16.23%	11.70%	10.53%
Social Studies Proficient in 2010	4.08%	5.94%	4.68%	7.60%
ELPA Scale Score in 2010	608.9	617.1**	607.4	607.7
Social Studies Scale Score in 2010	896.1	896.6	895.9	896.4

	Before Matching		After Matching	
	CBT (N=196)	PPT (N=4,010)	CBT (N=4,010)	PPT (N=196)
School Level Variables				
Number of ELL	26.21	93.42*** [a]	28.05	27.94 [a]
Number of Computers with High Speed Internet per ELL	41.56	24.80*** [a]	40.55	36.78
Average of Mean ELPA Scale Score in Last Three Years	627.6	629.9 [a]	625.5	624.2
Average of Mean Social Studies Scale Score in Last Three Years	896.7	896.8 [a]	896.2	896.0
Average of Proportion of Proficient on ELPA in Last Three Years	22.22	19.01** [a]	20.56	20.92
Average of Proportion of Proficient on Social Studies in Last Three Years	8.01	7.80[a]	8.25	7.74

Note. CBT=computer-based test. PPT=paper-based test. * $p<.05$, ** $p<.01$, *** $p<.001$.
[a] Variance significantly different at $p<.05$.

A.8: Grade 10

	Before Matching		After Matching	
	CBT (N=152)	PPT (N=3,532)	CBT (N=124)	PPT (N=124)
Individual Level Variables				
Native Language: Spanish	13.82%	34.74%***	8.06%	10.48%
Native Language: Arabic	4.61%	22.20%***	5.65%	4.03%
Economically Disadvantaged	61.18%	76.98%***	55.65%	51.61%
Disabled	5.92%	9.51%	6.45%	10.48%
Female	44.74%	45.22%	44.35%	43.55%
Ethnicity: Black	3.29%	4.28%	4.03%	4.03%
Ethnicity: Hispanic	36.18%	39.61%	22.58%	25.81%
Ethnicity: White	32.24%	38.05%	39.52%	41.13%
Ethnicity: Asian	27.63%	17.36%**	33.06%	28.23%
ELPA Proficient in 2010	14.47%	14.69%	12.10%	12.90%
ELPA Scale Score in 2010	628.6	626.9	626.4	627.9
School Level Variables				
Number of ELL	15.14	66.23*** [a]	15.75	14.94 [a]
Number of Computers with High Speed Internet per ELL	51.51	27.45*** [a]	47.54	44.60
Average of Mean ELPA Scale Score in Last Three Years	638.1	637.6 [a]	637.4	637.3
Average of Proportion of Proficient on ELPA in Last Three Years	22.92	24.56 [a]	22.77	23.30

Note. CBT=computer-based test. PPT=paper-based test. * $p<.05$, ** $p<.01$, *** $p<.001$.
[a] Variance significantly different at $p<.05$.

A.9: Grade 11

	Before Matching		After Matching	
	CBT (N=139)	PPT (N=2,593)	CBT (N=119)	PPT (N=119)
Individual Level Variables				
Native Language: Spanish	22.30%	33.13%**	19.33%	20.17%
Native Language: Arabic	4.32%	24.95%***	5.04%	6.72%
Economically Disadvantaged	71.94%	77.13%	73.11%	66.39%
Disabled	3.60%	9.10%*	4.20%	5.88%
Female	48.20%	46.12%	47.06%	44.54%
Ethnicity: Black	2.88%	4.74%	3.36%	2.52%
Ethnicity: Hispanic	35.97%	37.37%	33.61%	38.66%
Ethnicity: White	38.13%	41.73%	40.34%	32.77%
Ethnicity: Asian	22.30%	15.50%*	21.85%	25.21%
ELPA Proficient in 2010	13.67%	20.32%	13.45%	14.29%
Reading Proficient in 2011	7.91%	9.72%	8.40%	6.72%
Writing Proficient in 2011	8.63%	9.14%	7.56%	7.56%
Math Proficient in 2011	15.11%	6.02%***	13.45%	10.92%
Science Proficient in 2011	4.32%	2.20%	4.20%	0.84%
Social Studies Proficient in 2011	9.35%	8.14%	10.08%	5.04%
ELPA Scale Score in 2010	629.3	635.1*	628.1	630.5
Reading Scale Score in 2011	1061.8	1069.7*	1063.4	1065.2
Writing Scale Score in 2011	1048.6	1057.2* [a]	1048.1	1051.6
Math Scale Score in 2011	1065.7	1060.6	1064.9	1066.5
Science Scale Score in 2011	1069.6	1067.1	1069.1	1069.1
Social Studies Scale Score in 2011	1101.0	1099.8	1100.4	1099.9

	Before Matching		After Matching	
	CBT (N=2,593)	PPT (N=139)	CBT (N=119)	PPT (N=119)
School Level Variables				
Number of ELL	17.91	56.89*** [a]	18.41	15.87
Number of Computers with High Speed Internet per ELL	53.30	35.14* [a]	54.29	64.31
Average of Mean ELPA Scale Score in Last Three Years	645.8	643.4*** [a]	645.7	645.9 [a]
Average of Mean Reading Scale Score in Last Three Years	1070.0	1072.0* [a]	1069.5	1069.9
Average of Mean Writing Scale Score in Last Three Years	1055.6	1060.4*** [a]	1054.7	1056.3
Average of Mean Math Scale Score in Last Three Years	1073.5	1067.1***	1072.4	1073.6
Average of Mean Science Scale Score in Last Three Years	1073.4	1067.7*** [a]	1072.4	1073.5
Average of Mean Social Studies Scale Score in Last Three Years	1104.1	1100.7*** [a]	1103.3	1103.5 [a]
Average of Proportion of Proficient on ELPA in Last Three Years	27.41	27.18 [a]	28.64	29.13
Average of Proportion of Proficient on Reading in Last Three Years	12.14	12.29 [a]	11.80	12.55
Average of Proportion of Proficient on Writing in Last Three Years	10.72	10.08 [a]	10.16	9.63
Average of Proportion of Proficient on Math in Last Three Years	15.68	6.50*** [a]	14.42	13.51
Average of Proportion of Proficient on Science in Last Three Years	6.43	2.61*** [a]	5.86	5.22
Average of Proportion of Proficient on Social Studies in Last Three Years	13.76	8.77***	12.97	12.24

Note. CBT=computer-based test. PPT=paper-based test. * $p<.05$, ** $p<.01$, *** $p<.001$.

[a] Variance significantly different at $p<.05$.

A.10: Grade 12

	Before Matching		After Matching	
	CBT (N=115)	PPT (N=2,072)	CBT (N=106)	PPT (N=106)
Individual Level Variables				
Native Language: Spanish	14.78%	33.69%***	16.04%	21.70%
Native Language: Arabic	15.65%	22.78%	16.04%	13.21%
Economically Disadvantaged	59.13%	70.80%**	62.26%	59.43%
Disabled	5.22%	9.89%	5.66%	7.55%
Female	51.30%	50.72%	51.89%	48.11%
Ethnicity: Black	4.35%	4.30%	4.72%	6.60%
Ethnicity: Hispanic	30.43%	36.78%	33.02%	43.40%
Ethnicity: White	40.87%	40.93%	39.62%	34.91%
Ethnicity: Asian	24.35%	16.84%*	22.64%	15.09%
ELPA Proficient in 2010	9.57%	20.08%**	10.38%	13.21%
Reading Proficient in 2010	3.48%	10.23%*	3.77%	5.66%
Writing Proficient in 2010	0.87%	6.23%*	0.94%	0.94%
Math Proficient in 2010	5.22%	3.67%	4.72%	1.89%
Science Proficient in 2010	0.87%	2.08%	0.94%	0%
Social Studies Proficient in 2010	4.35%	6.66%	4.72%	7.55%
ELPA Scale Score in 2010	638.4	640.2 [a]	638.9	642.6
Reading Scale Score in 2010	1063.2	1067.3 [a]	1063.8	1069.4
Writing Scale Score in 2010	1043.9	1052.1	1044.5	1050.6
Math Scale Score in 2010	1066.3	1064.0	1064.8	1065.6
Science Scale Score in 2010	1057.0	1060.5	1056.7	1059.9
Social Studies Scale Score in 2010	1096.0	1099.3* [a]	1096.4	1098.8

	Before Matching		After Matching	
	CBT (N=2,072)	PPT (N=115)	CBT (N=106)	PPT (N=106)
School Level Variables				
Number of ELL	10.87	43.31*** [a]	10.63	10.56 [a]
Number of Computers with High Speed Internet per ELL	68.30	38.48*** [a]	69.29	66.18
Average of Mean ELPA Scale Score in Last Three Years	648.6	644.7* [a]	648.7	648.1
Average of Mean Reading Scale Score in Last Three Years	1076.5	1063.5*** [a]	1074.1	1076.5
Average of Mean Writing Scale Score in Last Three Years	1051.7	1047.7 [a]	1050.1	1052.5
Average of Mean Math Scale Score in Last Three Years	1080.3	1065.8***	1078.1	1074.3
Average of Mean Science Scale Score in Last Three Years	1073.7	1056.3*** [a]	1070.9	1068.7
Average of Mean Social Studies Scale Score in Last Three Years	1107.0	1099.5*** [a]	1105.4	1105.1
Average of Proportion of Proficient on ELPA in Last Three Years	22.08	21.23 [a]	21.58	19.46 [a]
Average of Proportion of Proficient on Reading in Last Three Years	10.87	8.30 [a]	8.02	6.06
Average of Proportion of Proficient on Writing in Last Three Years	4.87	5.59	4.34	2.95
Average of Proportion of Proficient on Math in Last Three Years	16.24	5.43*** [a]	13.85	12.82
Average of Proportion of Proficient on Science in Last Three Years	7.83	1.23*** [a]	4.72	3.76
Average of Proportion of Proficient on Social Studies in Last Three Years	21.62	6.78*** [a]	18.74	18.10 [a]

Note. CBT=computer-based test. PPT=paper-based test. * $p<.05$, ** $p<.01$, *** $p<.001$.

[a] Variance significantly different at $p<.05$.