PARCC

Mode Comparability Study based on Spring 2015 Operational Test Data

Junhui Liu, Terran Brown, Jianshen Chen, Usama Ali, Likun Hou and Kate Costanzo

Educational Testing Service

February 22, 2016

Table of Contents

List of Tables

List of Figures

Executive Summary

The Partnership for Assessment of Readiness for College and Careers (PARCC) is a state-led consortium working to develop next-generation assessments that more accurately, compared to previous assessments, measure student progress toward college and career readiness. The PARCC assessments include both English Language Arts/Literacy (ELA/L) and mathematics assessments in grades 3 to 8 and high school. Although the long-term goal of the PARCC assessment system is for digital administration, the initial test rollout supported both computer-based testing (CBT) and paper-based testing (PBT) modes of administration. One of the goals of the assessment was to report comparable scale scores across modes (CBT and PBT).

The mode comparability study was conducted to address the following two questions:

1. Is the construct invariant between the two modes of test administration?
2. Given that the construct remains the same, is student performance (e.g., mean, median, various quartiles) similar between the two modes?

To address these two research questions, a series of analyses were conducted using data from the spring 2015 operational tests of mathematics grades 5 and 7, Algebra I, Geometry, Algebra II, and ELA/L grades 3, 7 and 9. School districts selected the test administration mode, therefore the resulting CBT and PBT test-taker groups are not randomly equivalent. To make the groups more comparable, within each PARCC state, schools were matched on student background characteristics. Propensity score matching (PSM) was employed and was restricted to demographic information because only one participating PARCC state provided prior state test achievement data. The demographic characteristics used for matching included: ethnicity, gender, economic disadvantage status (EDS), disability conditions, and English learner (EL) status. Even with efforts to make groups comparable in terms of demographics, some grade levels still had differences in ethnicity and EDS due to the significantly different distribution of these covariates across modes within some states.

The following analyses were conducted for this mode comparability study:

i. Z-score comparisons (Section 3.2) to evaluate the similarity of item performance of the common items across modes.
ii. Differential item functioning (DIF; Section 3.3) to identify common items with differences in performance once test takers are matched on ability.
iii. Comparison of IRT item parameter estimates (Sections 5 and 7) to evaluate the similarity of item difficulty estimates and item discrimination parameter estimates based on separate within-mode IRT calibrations.
iv. Summary test statistics (Section 6) to compare “test-level” mean performance across modes. This analysis included effect sizes to determine the magnitude of possible mode effects.

The item level analyses showed that the differences in item difficulties were small for the majority of items. However, the Prose Constructed Response (PCR) trait items in ELA/L had larger differences in item difficulties compared to other item types; all differences favored PBT. The difficulties of the common items between modes were strongly correlated in nearly all subjects and grade levels indicating coherence in measuring the same construct. Although a very small percentage of items was

identified as having substantial differences across the two modes after accounting for test taker ability, many items were flagged for moderate differences across the two modes favoring PBT for ELA/L grades 3, 7 and 9 as well as for the Geometry test; the majority of these items in ELA/L were PCR trait items.

The test level analyses indicated that the IRT difficulty and discrimination parameters estimated separately within mode were highly correlated. The overall reliabilities based on total test raw scores and common item total raw scores were similar across modes. Common item total raw score effect sizes, an indicator of the magnitude of group differences, varied across subjects and grades in terms of magnitude and direction; therefore student performance on the common items varied across subjects and grades. The mean scale scores and effect sizes also varied across subjects and grades. In general the scale score effect sizes were similar to the performance of the common items, except for Algebra II. Plots graphing the probability of achieving each possible raw score on the common items in the two modes were evaluated. Overall, the differences were small and would not result in reported score differences across modes. However, for ELA/L grade 9 and Geometry differences that would affect reported scores were found in regions of the theta scale where large percentages of students were located.

Additional analyses were conducted on student data from the sole state (State S) that provided prior state assessment scores. Prior achievement data were used for adjustment to make the CBT and PBT groups more comparable. The scale score differences were largely reduced for mathematics grade 5, 7 and Algebra I after using the prior achievement data and the scale scores were generally comparable across modes for these tests. However, for other grades, particularly ELA/L grade 9 and Geometry, there were substantial differences in scores across mode.

There are several limitations to this study. First, there was no random assignment of students to testing mode. As previously noted, schools/districts selected the testing mode. Second, the effect sizes corresponding to the analysis of item difficulties and raw and scale score comparisons based on samples from the matching procedure are likely confounded with students’ ability differences across modes. Lastly, only one state provided the previous year’s state testing results. Therefore, the analyses involving State S may not generalize to other states.

Section 1: Overview of Mode Comparability Study for the PARCC Assessments

1.1 Introduction

The Partnership for Assessment of Readiness for College and Careers (PARCC) is a state-led consortium working to develop next-generation assessments that more accurately measure student progress toward college and career readiness than previous assessments. The PARCC assessments were aligned to the Common Core State Standards (CCSS) and were administered operationally beginning in the 2014-2015 academic year. The PARCC assessments include both English Language Arts/Literacy (ELA/L) and mathematics assessments in grades 3 to 8 and high school.

In 2015, the PARCC operational assessments comprised two components that contributed to a full summative (FS) score: a Performance-Based Assessment (PBA) administered after 75% of the academic year and an End-of-Year Assessment (EOY) administered after 90% of the academic year. Although the long-term goal of the PARCC assessment system is for digital administration, the initial test rollout supported both computer-based testing (CBT) and paper-based testing (PBT) modes of administration. The goal of the assessment system was to report comparable scale scores across modes (CBT and PBT).

According to the *Standard for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), whenever a test is administered on both computer and paper modes comparability studies must be conducted to support claims that test scores earned in either format may be used interchangeably and have the same interpretation. In preparation for this assessment, during the 2014 field test (FT) administration, PARCC commissioned a mode comparability research study that was designed to examine the items and scores across modes. The goal of the 2014 FT PARCC mode comparability study was to inform operational calibration, scaling, and test development decisions as the PARCC assessment moved to the operational phase.

The results of the 2014 FT mode comparability study (Brown, Chen, Ali, Costanzo, Chung, and Ling, 2015) indicated that a small mode effect may have existed in favor of PBT group. It was recommended that a similar study be repeated based on data collected during the first operational administration. PARCC commissioned the current mode comparability research study to evaluate to what degree scores from CBT and PBT form versions are comparable. The purpose of this document is to present background information, methods, results and a discussion of the mode comparability study based on operational data for the initial operational administration of the PARCC assessments.

1.2 Prior Research on Mode Comparability

Since the mid 1990’s there has been an expansion of computer technology to support teaching and learning in K-12 schools. A natural extension of increased computer usage in classrooms was to begin assessing students using this technology as it would be consistent with teaching and learning practices. Moreover, researchers and practitioners noted several benefits of administering assessments online, such as (a) flexible scheduling, (b) more efficient test administration, (c) increased test security, (d) quicker score reporting, (e) expanded range of content coverage, (f) use of innovative technology and item types, and (g) the integration of mixed media (e.g., movies, text, and audio clips; Bennett, 2003; Paek, 2005; Randall, Sireci, Li, & Kaira; 2012; Wan, Keng, McClarty & Davis, 2009). As such, states began developing computer-based tests as part of their assessment programs (Bennett, 2002 & 2003; Olson, 2003). However, integration of computers and online technology into K-12 classrooms as well as efforts to upgrade computer infrastructure (e.g., internet bandwidth and number of computers) has been uneven across schools, districts, and states. This has limited the ability of state assessment programs to completely transition away from paper-based tests. As states began supporting both modes of administration, there was a need for empirical research to support test score comparability.

Several mode comparability studies using state testing data have been conducted, including Kansas (Poggio, Glasnapp, Yang, Beauchamp, & Dunham, 2005), Texas (Keng, McClarty, & Davis, 2008; Way, Davis, & Fitzpatrick, 2006), Oregon (Oregon Department of Education, 2007), Maryland (Pearson Educational Measurement, 2010), Minnesota (Pearson Educational Measurement, 2012), and North Carolina (Lottridge, Nicewander, & Mitzel, 2010). These studies were conducted at the test level (e.g., comparing the mean total scores across modes, test reliability estimates, classification consistency, and confirmatory factor analyses) and at the item level (e.g., item statistics, item response theory parameter estimates, differential item functioning, Hierarchical Linear Modeling, and analysis of covariance).

The researchers conducting these mode comparability studies were trying to ascertain the comparability of the test items and measured student achievement across the two testing modes. The results of the mode studies have been mixed, with some studies finding no differences across modes (Kim & Huynh, 2007) and while other studies did find evidence of differences due to mode of administration, or mode effects (Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan 2008; Johnson & Green, 2006; Poggio, Glasnapp, Yang, & Poggio, 2005; Way, Lin & Kong., 2008). When mode effects were found, the differences across the two modes tended to be small. Findings of mode effects were also inconsistent across grade levels and subjects. A few studies showed that lower grade students were more likely to have difficulty answering items when presented on a computer than when presented on paper (Choi & Tinkler, 2002; Coon, McLeod, & Thissen, 2002). Way, Davis, and Fitzpatrick's (2006) investigation of Texas statewide tests showed that mode effects were more evident for ELA tests than other subjects.

Previous studies also found that mode effects were more evident for certain types of items that may require extra effort from students testing in one of the modes. Russell, Goldberg and O'Connor (2003) reviewed earlier literature in computer-based testing and found a number of factors that may influence the validity of computer-based tests, including the need to transfer problems from the screen to scratch workspace, item layouts, presentation of graphics, and the ability of students to work in the given mode. A series of studies have suggested modal differences on reading or history items with long passages and mathematics items involving graphing (Keng, et al. 2008; Way et al., 2006) and items that require the use of scrolling (Pommerich, 2004). Keng et al. (2008) and Way, et al. (2008) found differences in favor of PBT for items with long reading passages, mathematics items focusing on geometric relationships or requiring spatial reasoning. Johnson and Green (2006) noted that some students did not show their work for the CBT version of the item. Johnson and Green suggested that the show-your-work items may have more demands of memory and attention because online test takers would have to transfer the item to scratch paper, work out the answer, and then enter their response back into the CBT interface. Johnson and Green suggested that switching between the computer screen to read questions and scratch paper to work out answers had an impact on the online test takers – the paper test takers could work on their response directly on the test item, while the online test takers had to use scratch paper.

Finally, in reviewing previous research it was unclear whether these mode comparability studies examined comparability of paper tests placed into an electronic format or whether the CBT tests involve utilizing technology enhanced item types, or leveraging computer capacity to assess students differently. The extent to which the CBT version of the items were technology enhanced and changed the presentation and process of the test across modes may be the main reasons for mode differences suggested by previous literature (Pommerich, 2004).

1.3 Data Collection and Analyses Options for a Mode Comparability Study

According to Wan et al, (2009) there are two important steps to complete a mode comparability study: 1) collection of data for making score comparisons and 2) selection of the most appropriate analysis methods. The options for data collection and analysis methods are described below.

1.3.1 Data Collection Options

Three commonly used data collection designs are *common person*, *randomly equivalent groups*, and *quasi-experimental* (Wan et al., 2009). Each approach is briefly described below.

In the ***common person*** design, the same persons take the exam twice, once on computer and once on paper. Counter-balancing [i.e., some take the computer-based test first and others take the paper-based test first] is typically utilized to minimize potential practice or fatigue effects. The advantages of the *common person* design are the small sample sizes needed and that it is a very powerful method for detecting differences (Wan et al., 2009). The main disadvantage of this design is that test takers take the test twice which can lead to a lack of motivation due to fatigue or other factors.

The ***randomly equivalent groups*** design, if carried out properly, is the best approach, given practical issues. It involves a random assignment of test takers to either the CBT or the PBT groups. "The advantages of this design are that test takers only need to test once. Additionally, since the two groups are, in theory, the same on all important characteristics, no further manipulation of the groups is necessary" (Wan et al., 2009, p. 1). The main disadvantage of this design is that it is often not practical to randomly assign test takers in an operational administration.

In the ***quasi-experimental*** design, two existing groups are administered tests in different modes and the results are compared. The two existing modes could be delivered to all students in a classroom, or a school, or a district and resulting groups are *not* by nature randomly equivalent. This design may require additional manipulation of the data at the time analyses are completed in an attempt to make the CBT and PBT groups equivalent (e.g., propensity score matching; Rosenbaum & Rubin, 1983). In short, a "quasi-experimental design poses minimal burden on institutions conducting data collection and could easily be part of the regular testing administration. However, this design requires additional demographic information about each student, and the quality of the study results is dependent on the degree of similarity of the samples created" (Wan et al., 2009, p. 2).

For the year one operational mode comparability study, *common person* and *randomly equivalent group* designs were not viable options. Districts and/or schools determined which students would test online and on paper and random assignment was not applied across modes. As a result, it could not be assumed that students administered the tests online were equivalent to students administered the tests on paper in terms of their underlying English language arts or mathematics abilities. For the year one operational mode comparability study, the only option was to use a *quasi-experimental* approach, where demographic characteristics of the CBT and PBT samples were used to adjust samples so they were more similar across modes.

The PARCC technical advisory committee recommended adjusting the samples as needed by matching students across modes based on their state assessment scores from previous assessments. To do this it would have been necessary to obtain scores for students based on existing state tests. Unfortunately, only one state provided previous test data at the time the analyses were conducted.

1.3.2 Appropriate Analysis Methods

When conducting a mode comparability study, the study should address the following two questions:[1]

1. Is the construct invariant between the two modes of test administration?

2. Given that the construct remains the same, is student performance (such as mean, median, various quartiles) similar between the two modes?

To address the first question, the following analyses are appropriate:

i. *Z-score comparisons (Section 3.2) – This analysis is designed to evaluate the similarity of item performance of the common items across modes based on classical test theory methods. Lack of consistency across modes might indicate the item is measuring a different attribute and might warrant further inspection.*

ii. *Differential item functioning (DIF) analysis (Section 3.3) – This analysis looks at differences in the performance on common items across modes of test takers matched on ability. It is important to look for patterns (e.g., specific item types) and investigate further, items that may show large levels of significant DIF.*

iii. *Confirmatory factor analyses (CFA) – This analysis specifies a unidimensional and/or multidimensional factor structure and tests whether the underlying structure is the same or consistent across modes. If the factor structure is not consistent, then the tests may not be measuring the same construct.*

iv. *Analyses of IRT Item Parameter Estimates (Section 5&7) – When IRT calibration and scaling procedures are used for linking parameter estimates across modes, it is important to evaluate whether the estimated difficulty and discrimination parameter estimates based on separate within-mode IRT calibrations measure the same construct. High, positive correlations between common items across modes are expected. Calibration of common item parameters should not be severely impacted by unique items in each mode. Strong correlations of common item parameters estimated in the presence and in the absence of unique items provide evidence of the same construct being measured across CBT and PBT forms. Additionally, if the common items perform similarly across modes, then the test characteristic curve (TCC) for items appearing on paper should look nearly identical to the TCC for items appearing on computer, after scaling.*

[1] A third, very difficult to answer, question is: Does the relationship between modes estimated by this study likely generalize to other test material and/or student groups? While additional research would be required to answer this question, consideration of this factor serves as an important caveat.

To address the second question, the following analyses are appropriate:

i. *Summary Test statistics (Section 6) – If there appears to be construct invariance and there are either randomly equivalent or matched samples, it is appropriate to compare and summarize "test-level" mean performance across groups. The success of the propensity score matching is critical to the validity of such analyses.*

ii. *Effect sizes (Section 6) – Mean differences at the "test-level" can be converted into effect sizes to determine the magnitude of any possible mode effects.*

All of the analysis methods listed above, except CFA, were used in this mode comparability study and are described in the appropriate section of the report with the results (i.e., in Sections 3 through 7). CFA was not repeated for the operational mode comparability study because there were no "sister forms" (i.e., CBT and PBT forms comprising mostly the same items) and significant variation on the number of common items across operational CBT and PBT forms. Conclusions about construct invariance from CFA conducted on a particular form cannot be generalized to other forms.

1.4 Lessons Learned from the 2014 Field Test

A number of analyses and special studies were conducted using 2014 field test (FT) data to inform decisions related to the operational calibration and scaling in 2015. The goal of the 2014 FT PARCC mode comparability study (Brown et al., 2015) was to evaluate to what extent scores from the CBT and PBT form versions could be considered comparable. The findings indicated that scores from the FT forms were not comparable across modes in a strict sense, particularly for PBA. However, there was substantial evidence indicating that the differences in comparability across modes were relatively minor. When comparing the performance of the common items, there were small effect sizes in favor of PBT for the mathematics and ELA/L PBA assessments and negligible effect sizes for EOY and full summative assessments.

Specifically, the DIF results indicated that a small number of items in the ELA/L (i.e., 0 to 7 items per grade) item pool and a slightly higher number of items in the mathematics (i.e., 2 to 17 items per grade) item pool possessed a substantial degree of differences across modes.

There were two implications for the operational calibration and scaling plan based on these findings. First, since DIF clearly existed for some items in the 2014 FT study, it was appropriate to calibrate operational CBT and PBT items separately for each grade/subject. Second, when scaling PBT item parameter estimates to the CBT scales, the exclusion rules used for linking 2014 FT items was appropriate and should also be used for the 2015 operational administration (i.e., items flagged for positive and negative C-DIF should be removed from the linking sets). Common items that behave differently across modes should be treated as separate unique items and make use of both CBT and PBT item parameter estimates for generating operational conversion tables.

1.4.1 Limitations of 2014 FT Study

The following factors may have limited the conclusions of the 2014 FT mode comparability study:

1) Student motivation to perform their best was likely low. (This is often an issue for standalone field tests because there are no incentives/consequences for students.)

2) The degree of implementation of the CCSS was different across states.

3) Many of the item types (e.g., technology-enhanced items) on the field tests were being seen by students for the first time and, therefore, may have been novel.

It was concluded that it will take time for the CCSS to be implemented and for students to become accustomed to the new item types. The field test analyses were designed to inform the operational administration, and should be considered preliminary.

1.4.2 Recommendations Based on the 2014 FT Results

The following were three recommendations for the 2015 operational calibration and scaling procedures based on the 2014 FT results:

1) All 2014 FT study results should be considered preliminary because the data were based on a standalone field test that was: a) administered to students who may have been unmotivated; b) based on CCSS content that was either not fully implemented or implemented differently across states; and, c) used item types that may have been novel to students.

2) The mode comparability study should be repeated with year one operational data to confirm the 2014 field test findings.

3) The field test results support the calibration of PBA and EOY items together, and the calibration of CBT and PBT data separately. It seemed appropriate to proceed with the operational analyses using this approach.

Again, it was recommended that the 2014 FT studies be repeated using 2015 operational data because changes were expected in student performance with respect to motivation and exposure to CCSS implementation. In addition, sample sizes would be substantially larger for the operational administration compared to the FT administration. Additionally, the planned operational equating procedures were informed by the FT results. Therefore, if the repeated study results, based on operational data, were different from the FT results, then adjustments to the equating procedures in future years may be needed. Repeating the mode comparability study on operational data will help inform if reported scores are comparable across forms, modes, and devices.

1.5 Mode Comparability Study Limitations

1.5.1 Data Collection Limitations

A major limitation of the 2015 PARCC operational mode comparability study relates to the data collection design. Test takers who took the PBA and EOY components on a *computer* were randomly assigned to one of dozens of CBT form combinations. Test takers who took the PBA and EOY components on *paper* were also randomly assigned to one of a dozen or so PBT form combinations. However, there was no random assignment across administration modes (CBT and PBT). Schools/districts decided their students' testing mode. The resulting student samples in different testing modes were likely not randomly equivalent. In order to conduct mode comparability analyses on the operational data, it was necessary to make additional adjustments to the samples (e.g., propensity score matching; Rosenbaum & Rubin, 1983).

This *quasi-experimental* approach relied on demographic characteristics of the CBT and PBT groups to adjust samples so they were more similar in ability across modes. The quality of the propensity score matching depends on the quality of the demographic data available for matching. Ideally, mode comparability samples would be adjusted as needed to be matched on individual student test scores on previous state assessments. Due to the unavailability of state assessment data at the time of analyses, however, the current mode study only included demographic variables when matching CBT and PBT students. A number of demographic characteristics were used for matching including: ethnicity, gender EDS, EL background. Even with efforts to make groups comparable in terms of demographics, some grade levels still had differences in ethnicity and EDS due to the significantly different distribution of these covariates across modes within some states.

1.5.2 Variation in the Number of Common Items across Modes

In the 2014 FT design, an effort was made to build what were called “sister forms” across a subset of CBT and PBT forms. Sister forms were constructed to be primarily the same items on both CBT and PBT versions, except for technology-enhanced items that could not be delivered on paper. The sister forms were used to conduct many of the mode comparability analyses, especially the CFA analyses. However, in the operational setting it was much more difficult to create full summative sister forms. PBA and EOY forms were spiraled randomly within-mode so that there were approximately 64 CBT versions and 16 PBT operational core versions for some grades and subjects. This very complex spiraling design in some cases placed limitations on the form-level analyses that could be carried out.

1.6 Overview of the Report

Section 1, serves as an introduction to the report and includes sections related to data collection designs, appropriate analyses, results from the 2014 FT mode comparability study, the research questions of interest, and study limitations. Section 2 describes the 2015 PARCC operational test design and the numbers of common items across administration modes. Section 3 summarizes the methods and results of propensity score matching of CBT and PBT data. The analysis methods and results are presented in Sections 4 through 7. Section 8 discusses the implications and limitations of current study and provides suggestions for future mode comparability study.

Section 2: Operational Test Design

2.1 Overview

The PARCC Spring 2015 operational test includes nine ELA/L tests (grades 3 to 11) and 12 mathematics tests (grades 3 to 8, and six end-of-course [EOC] tests at the high school level – Algebra I, Geometry, Algebra II, Mathematics I, Mathematics II, and Mathematics III).

All test forms were constructed to match operational test blueprints in terms of content, item types, and test length. The PARCC assessment design entailed two components. The performance based assessment (PBA) was administered after approximately 75 percent of instruction in a school year has occurred, and the end-of-year (EOY) assessment was administered after approximately 90 percent of instruction has occurred. Together, the PBA and EOY components composed the full summative (FS) operational assessment. Paper-based tests (PBT) and computer-based tests (CBT) were available for both ELA/L and mathematics. Within each of the assessment components, schools/districts determined the mode of administration for their students. For the majority of schools, there was one mode of administration across test components for all students except for those who required special accommodations.

The number of operational core forms for each grade/subject is presented in Table 2.1 for ELA/L and mathematics. Operational test forms include embedded field test items. Test forms within a grade, component, and content area are spiraled at the student level to support the distribution of field test sets across randomly equivalent groups.

Table 2.1 Number of Core Operational Forms per Grade/Subject for Each Component and Administration Mode for ELA/L and Mathematics

	ELA/L				Mathematics			
	CBT		PBT		CBT		PBT	
Grade/ Subject	PBA	EOY	PBA	EOY	PBA	EOY	PBA	EOY
Grade 3	6	6	4	4	6	6	4	4
Grade 4	6	6	4	4	6	6	4	4
Grade 5	6	6	4	4	6	6	4	4
Grade 6	6	6	4	4	6	6	4	4
Grade 7	6	6	4	4	6	6	4	4
Grade 8	6	6	4	4	6	6	4	4
Grade 9	8	8	4	4				
Grade 10	8	8	4	4				
Grade 11	6	6	4	4				
Algebra I					8	8	4	4
Geometry					8	9*	4	4
Algebra II					6	6	4	4
Integrated Mathematics I					2	2	2	2
Integrated Mathematics II					2	2	2	2
Integrated Mathematics III					2	2	1	2

Note: *For Geometry CBT EOY there were 9 core forms, instead of 8, because one item is different on two versions.

2.1.1 CBT and PBT Form Construction

During the test development process, steps were made to ensure that PBT and CBT items were comparable. The development process began by evaluating the test blueprint and identifying blueprints of items that could be assessed on PBT test forms. For mathematics, the goal was verify that 50 percent of each evidence statement could be assessable on a PBT format. The development process for PBT items started with looking at each technology-enhanced item that need a replacement. The construct of the original item along with its cognitive complexity was examined prior to developing the PBT item. In instances where the same construct could not be maintained between the technology-enhanced item and the PBT replacement, alternative decisions were made. First, efforts were made to find another evidence statement within the same probability cluster with a similar construct. If this was unsuccessful, another construct using the same Evidence Statement was developed.

Since the PARCC ELA/L assessments report at the claim and subclaim levels, a decision was made to create replacement evidence-based selective response (EBSR) items for the technology-enhanced constructed response (TECR) items at the same subclaim level. Although the replacement items could have the same evidence statement and measure the same content as the TECR items, this was not necessarily the case. The PBT items were always written to the same subclaim as the TECR items that needed replacement. Replacement items did not need to be the same complexity as the TECR item. Multimedia passages did not require replacement items for PBT test forms, since they will never be tested on a PBT form.

All PBT items went through committee review to confirm the appropriateness of the items. During later meetings, item review committees made recommendations for item sets for PBT forms.

2.1.2 Linking CBT and PBT Parameter Estimates

The 2-parameter logistic (2PL) was used for item parameter estimation, based on an evaluation of model fit in the field test results (PARCC, 2015). The generalized partial credit (GPC) model was also used.

Parameter estimates needed to be placed onto a common scale across administration modes for each grade/subject to maintain comparable operational scale scores. There were two potential approaches considered for doing this:

1) Perform a single **concurrent calibration** across components (EOY and PBA) and administration modes (CBT and PBT) for each grade/subject.

2) Perform **calibrations within administration mode**, and use Stocking and Lord to transform the PBT result to the CBT scale.

The technical literature on this distinction is equivocal; some papers (Beguin & Hanson, 2001; Beguin, Hanson, & Glas, 2000) suggest that the “concurrent calibration” approach performs better, and others (Hanson & Beguin, 2002; Kim & Kolen, 2007) suggest that the separate calibrations and a scale transformation found in the “calibrations within administration mode” approach performs better. Not having a clear reason to believe that one approach would provide more accurate results, the decision focused on secondary considerations.

The “concurrent calibration” approach is most efficient, but only provides marginally-useful fit statistics for evaluating differences in parameter estimates across administration modes. Should large fit statistics be traced back to different performance of common items in CBT and PBT administrations, a new calibration is required, treating those common items as unique items in the two administration modes.

The “calibrations within administration mode” approach is less efficient in that it initially requires two item calibrations. However, the scale transformation procedures provide an opportunity to directly evaluate differences in item characteristic curves for common items across CBT and PBT administrations. This approach also makes it straightforward to use different parameter estimates for an item that appears to be the same in CBT and PBT administration, but performs substantially differently. This approach was ultimately chosen because the field test administration identified items that would require different parameter estimates to treat CBT and PBT students fairly.

Separately calibrating PBA and EOY items was briefly considered because of the assumption being made that each student has a single ability despite having a month or more between the PBA and EOY administrations. This approach was discarded because the test form structure does not include common items across these administrations.

2.2 Mode Comparability Study Sample and Forms

Test takers participating in the 2015 PARCC operational administration were not randomly assigned to the paper and online administration modes and the resulting groups are not randomly equivalent. For the PARCC mode comparability study a quasi-experimental design was used along with propensity score matching (PSM; Rosenbaum & Rubin, 1983) to achieve pseudo-random equivalent groups (Refer to Section 3 for details about propensity score matching). The PSM analyses were conducted for a selected full summative form pair (i.e., CBT-PBT form pair) for a selected grade for each grade-level span. All mode comparability analyses, including item analysis (IA) and item response theory (IRT), were conducted on the matched samples except for differential item functioning (DIF). DIF was conducted for each content area and grade level on the 2015 PARCC operational analysis data. The selected graded levels and core form pairs are listed in Table 2.2 and the criteria of selecting form pairs were as follows:

- There were at least 30% of total test items and 30% of total test points in common between the CBT and PBT core forms;
- All PCR items were common items between ELA/L CBT and PBT core forms;
- There were a limited number of problematic items based on item analyses; and,
- There was a minimum target sample size of 1,200 valid cases per item/task.

Table 2.2 List of Selected Grade Levels and Form Pairs for 2015 PARCC Operational Test Mode Comparability Study for ELA/L and Mathematics

	EOY form		PBA form	
Subject/Grade	PBT core form	CBT core form	PBT core form	CBT core form
Mathematics 5	B	C	A	A
Mathematics 7	A	A	A	A
Algebra I	A	A	A	A
Geometry	C	E	A	A
Algebra II	A	A	A	A
ELA/L 3	A	A	A	A
ELA/L 7	A	A	A	A
ELA/L 9	B	B	A	A

Note: Letters, e.g. A, B, and C, indicate different core forms for a 2015 PARCC Operational Test.

2.3 Common Items across Modes

In response to several practical constraints, to meet the blueprints (e.g., inclusion of technology enhanced items in CBT forms), no one core form served as an equivalent test form in terms of all items and content being exactly the same between computer and paper administration modes at each grade level. When selecting core form pairs for the mode comparability study, priority was given to those with more common items and higher number of points of common items across the administration modes.

Tables 2.3 summarizes the number of common items between CBT and PBT forms for each subject/grade as administered during the operational test.

Table 2.3 Number of Common Items between CBT and PBT Forms

	EOY		PBA		Total Common Items			
Subject/ Grade	Number of Items	Number of Points	Number of Items	Number of Points	Number of Items	Number of Points	% of Items of Total Test	% of Points of Total Test
Mathematics 5	17	19	6	11	23	30	44	37
Mathematics 7	15	21	6	15	21	36	42	44
Algebra I	14	20	6	22	20	42	38	43
Geometry	17	27	13	35	30	62	57	64
Algebra II	11	16	8	28	19	44	37	44
ELA/L 3	9	18	22	70	31	88	84	88
ELA/L 7	20	40	27	91	47	131	96	97
ELA/L 9	19	38	25	87	44	125	88	91

Section 3: Propensity Score Matching

3.1 Overview

Propensity score matching (PSM) is a statistical matching technique that attempts to estimate the effect of a treatment, mode, or other interventions by accounting for the covariates that predict receiving the treatment. This statistical method attempts to reduce bias due to confounding variables that could be found in an estimate of the treatment or mode effect obtained from simply comparing outcomes among units that received the different treatment or in a different mode. As noted in Rosenbaum and Rubin (1983), propensity score analysis is a practical tool for reducing selection bias by establishing balance on observable covariates, where the propensity score is a scalar function of covariates so that subjects who match on their propensity scores can be treated as having similar covariate background.

3.2 Objectives

This section provides a description of the methodology and data specification of PSM analysis, as well as the evaluation of the matching results. The PSM analyses were conducted for each selected full summative form pair (i.e., CBT-PBT form pair) for the selected grade for each grade-level span. All mode comparability analyses, except for DIF, were conducted on the matched samples resulted from the PSM analysis. Details of the selected grade levels and core form pairs are listed in Table 2.2.

3.3 Method

In the spring 2015 PARCC administration, schools/districts determined the testing mode for their students. School-level propensity score matching was applied to the data to help adjust for the fact that students were not randomly assigned. Students who did not test in the same mode as the majority of their peers within a school were excluded from the analysis of propensity score matching. In some instances students did not test in the same mode for both components (PBA and EOY) of the test. Additionally, there were rare cases, in which students switched schools, districts, or states during the PARCC administration, which impacted their test mode in the process. Therefore students whose testing mode changed between the administrations of PBA and EOY were removed from the propensity score matching analysis. For each selected full summative form, the main testing mode for the majority of the students within that grade was determined for each school.

School-level PSM was conducted within each state that participated in the 2015 PARCC tests. This process is based on the assumption that schools are somewhat more similar within states than between states. One of the important assumptions for PSM methods is the strong ignorability assumption (Rosenbaum & Rubin, 1983), which presupposes that all covariates are related to mode selection and potential outcomes are included in the propensity score model. If PSM is conducted across states, the differences between states cannot be easily captured in the data. After matched subsamples for each state were obtained, the subsamples were combined into a single data pool for further analyses.

For each of the school-level datasets within state, PSM employed a predicted probability of testing mode selection, that is, CBT vs. PBT – based on observed background covariates, obtained from logistic regression to create a pseudo-equivalent group of schools. The “glm” function for fitting generalized linear models in R was used to conduct the logistic regression analysis and calculate propensity scores.

Selecting the covariates required to generate the propensity score can be complex. As Parsons (2001) noted, the propensity score is only as good as its model. Following a recommended approach, efforts were made to collect as much information as possible on students and then refine the logistic regression model focusing on the available variables that occurred prior to the testing mode selection, and were most likely to have influenced the CBT vs. PBT group participation. Initially, students' prior achievement in a state assessment of related subjects was proposed as one of the covariates in the model. However, due to the unavailability of state assessment data for a majority of the students at the time of analyses, the current mode comparability study only included demographic variables in the propensity score matching analysis. Only schools with complete information on the selected school-level covariates were included in the analysis of propensity score matching. The following covariates were included in the final logistic regression model and were hypothesized to be associated with testing mode selection by the schools:

- percentage of African American students,
- percentage of Hispanic students,
- percentage of White students,
- percentage of Female students,
- percentage of economically disadvantaged (EDS) students,
- percentage of English learner (EL) students, and
- percentage of students with disabilities (SWD).

Likelihood-based pseudo R^2 was calculated to evaluate the model goodness-of-fit of the final logistic regression model for the propensity score calculation. The summary of likelihood-based pseudo R^2 is listed in Table 3.1 and shows that, for some of the states, the final logistic regression was limited in terms of predicting the mode selection because the value of the pseudo R^2 was low. Logistic regression model uses maximum likelihood for estimation and therefore, ordinary least square (OLS) R^2 in regular regression does not exist. To evaluate the goodness-of-fit of logistic models, pseudo R^2s were often used. Even though "pseudo" R^2 looks like OLS R^2 in the sense that they are on a similar scale (ranging from 0 to 1) with higher values indicating better model fit, they cannot be interpreted as one would interpret an OLS R^2.

The most frequently used PSM methods are greedy match and optimal full match. The greedy match uses the nearest-neighbor approach; the first match that is within a minimum acceptable distance between propensity scores is chosen and maintained. The optimal full match looks for the closest distance between any matched combinations and thus reconsiders matches until the closest or optimal match is established. With optimal full match, each controlled unit can be matched while the overall discrepancy is minimized. Optimal full match is always as good as, and often better than, greedy match (Rosenbaum, 1989). The optimal full match can be combined with the use of propensity score calipers, where the closest match is picked in terms of Mahalanobis distance metric from a restricted subset or caliper of potential controls who were close to the treated unit on the propensity score. In this study, the optimal full match combined with the use of propensity score calipers was implemented with the "fullmatch" function in the "optmatch" package in R (Hansen & Klopfer, 2006).

To check the performance of propensity score matching (Austin, 2008; Harder, Stuart & Anthony, 2010), the standardized mean difference index (Cohen's *d*; Cohen, 1988) for each continuous covariate between the CBT group and the BT group at the school level was compared before and after matching. Specifically, Cohen's *d* is obtained by

$$Cohen's\ d = |\bar{x}_{CBT} - \bar{x}_{PBT}| / \sqrt{(s_{CBT}^2 + s_{PBT}^2)/2} \qquad (1)$$

where $\bar{x}_{CBT}$ and s_{CBT}^2 are the average value and variance of the examined covariate by school in the CBT group; $\bar{x}_{PBT}$ and s_{PBT}^2 are the average value and variance of the examined covariate by school in the PBT group. Cohen's *d* greater than 0.2 *SD* indicates lack of desirable covariate balance (Rubin, 2001; Shadish, Clark, & Steiner, 2008). Given all the factors that were considered for matching, it is quite difficult to get satisfactory covariates balanced on all variables of interest. Since social economic status tends to be more correlated with student achievement (Sirin, 2005), it was critical to achieve balance on percentage of economically disadvantaged students first. Additionally, a decision was made to accept aggregate covariance balance across all covariates based on average Cohen's d. Ideally the Cohen's d on all variables after matching should be less than 0.1. However, due to the substantial differences in CBT and PBT student distributions on matching variables, the criteria was adjusted in current study for the matching process to converge. The current study used to criteria to evaluate the balance of the covariates n the matched sample:

1) Cohen's *d* corresponding to the percentage of EDS students did not exceed 0.2, and

2) average Cohen's *d* across all selected covariates did not exceed 0.2.

3.4 Results

Table 3.1 lists the likelihood-based pseudo R^2 of the final logistic regression model in calculating propensity scores. Since logistic regression was run within each state for each subject/grade level, Table 3.1 provides a summary of pseudo R^2 of the logistic regressions across states. Overall the pseudo R^2 values were quite low indicating using demographic variables alone to predict testing mode might be inadequate.

Table 3.1 Summary of Pseudo R^2 of Logistic Regression in Propensity Score Matching

Subject/Grade	Minimum Pseudo R^2	Maximum Pseudo R^2	Median Pseudo R^2
Mathematics 5	0.021	0.307	0.111
Mathematics 7	0.025	0.421	0.286
Algebra I	0.008	0.193	0.149
Geometry	0.041	0.432	0.116
Algebra II	0.031	0.704	0.104
ELA/L 3	0.017	0.144	0.090
ELA/L 7	0.053	0.246	0.154
ELA/L 9	0.094	0.162	0.116

For each subject and grade included in this study, Tables 3.2 through 3.11 provide the comparison of descriptive statistics of the covariates used for matching (i.e., percentage of African American students, percentage of Hispanic students, percentage of White students, percentage of Female students, percentage of economically disadvantaged (EDS) students, percentage of English learner (EL) students, and percentage of students with disabilities (SWD)) in the CBT and PBT schools separately before and after matching. Also included in these tables are Cohen's *d* for each of the selected covariates as well as the average Cohen's *d* calculated on the matched sample for evaluating the performance of propensity score matching. As shown in Tables 3.2 through 3.11, all matched samples with all available states included met the covariance balance criteria in that Cohen's *d* for percentage of EDS students was less than 0.2 and average Cohen's *d* across all selected covariates was less than 0.2, except for the Mathematics grade 5 and ELA/L 3 form pairs.

As shown in Table 3.5, for the Mathematics grade 5 form pair, when the school-level sample included all available states, Cohen's *d* for percentage of EDS students after matching was 0.205, and the average Cohen's *d* across all selected covariates after matching was 0.163. After a series of investigations, it was determined that the distributions of ethnic groups and EDS students between CBT and PBT students in one of the states (identified as State A for confidentiality) were very different. Table 3.12 shows the descriptive statistics of the variables selected for matching for State A. Even in the matched sample, the averages for percentage of African American, percentage of White, and percentage of EDS were still significantly different between CBT and PBT groups for State A. After excluding State A schools from the overall sample, the matching results improved significantly. As shown in Table 3.6, for the matched sample without State A schools, Cohen's *d* for percentage of EDS students was 0.037, and the average Cohen's *d* across all selected covariates after matching was 0.115.

Similarly, shown in Table 3.8, for ELA/L 3 form pair, when the school-level sample included all available states, Cohen's *d* for percentage of EDS students after matching was 0.327, and the average Cohen's *d* across all selected covariates after matching was 0.175. The distribution differences across modes in State A again impacted the propensity score matching results. After excluding State A schools from the overall sample, the matching results improved significantly. As shown in Table 3.9, for the matched sample without State A schools, Cohen's *d* for percentage of EDS students was 0.089, and the average Cohen's *d* across all selected covariates after matching was 0.068. Table 3.13 shows the descriptive statistics of matching variables for State A for the ELA/L 3 form pair. Results show significant difference in the distributions of ethnic groups and EDS students between CBT and PBT groups for State A.

Table 3.2 Characteristics of Schools before and after Propensity Score Matching for Algebra I Form Pair

Variable	Statistics	Overall					
		Before Matching		After Matching		Before Matching Cohen's *d*	After Matching Cohen's *d*
		CBT	PBT	CBT	PBT		
School	*N*	2061	780	1185	591	-	-
% African American	Mean	15.75	16.94	14.82	14.51	0.038	0.010
% Hispanic	Mean	16.52	9.89	12.21	8.37	0.238	0.149
% White	Mean	55.34	50.42	59.61	53.89	0.117	0.134
% Female	Mean	49.35	47.22	47.24	48.43	0.060	0.032
% EDS	Mean	37.01	33.78	33.75	30.85	0.081	0.073
% EL	Mean	3.44	3.67	1.59	2.63	0.016	0.097
% SWD	Mean	7.52	24.59	6.22	15.73	0.581	0.372
Students	*N*	5339	7952	2638	5475	-	-
						0.162*	0.124*

Note: *Average Cohen's *d* of all covariates

Table 3.3 Characteristics of Schools before and after Propensity Score Matching for Algebra II Form Pair

Variable	Statistics	Overall					
		Before Matching		After Matching		Before Matching Cohen's *d*	After Matching Cohen's *d*
		CBT	PBT	CBT	PBT		
School	*N*	1013	134	427	108	-	-
% African American	Mean	13.75	12.30	8.05	10.52	0.053	0.109
% Hispanic	Mean	20.21	24.39	18.51	20.69	0.121	0.065
% White	Mean	54.25	48.26	66.08	56.11	0.145	0.247
% Female	Mean	49.34	43.42	48.26	46.48	0.177	0.050
% EDS	Mean	36.22	38.04	31.23	36.77	0.047	0.143
% EL	Mean	2.49	4.96	1.06	2.07	0.155	0.104
% SWD	Mean	5.73	12.98	3.19	6.46	0.277	0.168
Students	*N*	4054	1012	1279	893	-	-
						0.139*	0.127*

Note: *Average Cohen's *d* of all covariates

Table 3.4 Characteristics of Schools before and after Propensity Score Matching for Geometry Form Pair

Variable	Statistics	Overall					
		Before Matching		After Matching		Before Matching Cohen's *d*	After Matching Cohen's *d*
		CBT	PBT	CBT	PBT		
School	*N*	1000	274	452	227	-	-
% African American	Mean	8.46	9.79	7.28	8.55	0.053	0.051
% Hispanic	Mean	20.92	10.10	11.41	8.64	0.352	0.105
% White	Mean	57.54	53.98	66.63	57.18	0.082	0.217
% Female	Mean	47.67	52.81	43.91	52.35	0.134	0.204
% EDS	Mean	36.57	27.48	27.20	24.61	0.225	0.066
% EL	Mean	3.60	2.57	2.66	1.58	0.075	0.093
% SWD	Mean	6.97	4.26	2.95	2.01	0.148	0.077
Students	*N*	2510	1097	868	857	-	-
						0.153*	0.116*

Note: *Average Cohen's *d* of all covariates

Table 3.5 Characteristics of Schools before and after Propensity Score Matching for Mathematics Grade 5 Form Pair

Variable	Statistics	Overall					
		Before Matching		After Matching		Before Matching Cohen’s *d*	After Matching Cohen’s *d*
		CBT	PBT	CBT	PBT		
School	*N*	4208	2025	2492	1447	-	-
% African American	Mean	16.14	28.08	16.15	26.08	0.328	0.270
% Hispanic	Mean	19.68	16.88	14.97	15.73	0.086	0.025
% White	Mean	53.73	39.42	59.94	42.95	0.341	0.400
% Female	Mean	50.80	50.06	52.53	51.29	0.022	0.035
% EDS	Mean	43.43	53.92	42.75	51.47	0.250	0.205
% EL	Mean	3.18	5.02	2.12	2.95	0.128	0.073
% SWD	Mean	4.80	7.23	2.87	4.88	0.134	0.132
Students	*N*	10196	11196	5656	7530	-	-
						0.184*	0.163*

Note: *Average Cohen’s *d* of all covariates

Table 3.6 Characteristics of Schools before and after Propensity Score Matching for Mathematics Grade 5 Form Pair with State A Excluded

Variable	Statistics	Overall					
		Before Matching		After Matching		Before Matching Cohen’s *d*	After Matching Cohen’s *d*
		CBT	PBT	CBT	PBT		
School	*N*	3196	1317	1514	857	-	-
% African American	Mean	17.80	23.66	19.50	22.67	0.166	0.086
% Hispanic	Mean	19.66	11.11	11.99	8.08	0.286	0.153
% White	Mean	51.84	44.58	59.37	48.40	0.173	0.254
% Female	Mean	50.36	49.78	53.06	50.81	0.017	0.062
% EDS	Mean	42.21	44.82	39.57	41.12	0.063	0.037
% EL	Mean	3.45	5.22	2.18	2.52	0.114	0.028
% SWD	Mean	6.29	11.06	4.66	8.16	0.226	0.182
Students	*N*	7721	7520	3272	4572	-	-
						0.149*	0.115*

Note: *Average Cohen’s *d* of all covariates

Table 3.7 Characteristics of Schools before and after Propensity Score Matching for Mathematics Grade 7 Form Pair

Variable	Statistics	Overall					
		Before Matching		After Matching		Before Matching Cohen's *d*	After Matching Cohen's *d*
		CBT	PBT	CBT	PBT		
School	*N*	2753	1165	1195	594	-	-
% African American	Mean	17.65	24.48	18.18	23.00	0.199	0.133
% Hispanic	Mean	18.35	15.13	13.25	14.41	0.109	0.042
% White	Mean	54.83	42.40	58.76	47.41	0.306	0.267
% Female	Mean	50.23	42.18	45.50	42.71	0.256	0.080
% EDS	Mean	45.01	50.94	44.55	51.62	0.150	0.173
% EL	Mean	3.12	7.78	2.21	4.99	0.284	0.199
% SWD	Mean	6.47	25.53	5.16	11.21	0.708	0.278
Students	*N*	10628	18307	4010	7773	-	-
						0.288*	0.167*

Note: *Average Cohen's *d* of all covariates

Table 3.8 Characteristics of Schools before and after Propensity Score Matching for ELA/L Grade 3 Form Pair

Variable	Statistics	Overall					
		Before Matching		After Matching		Before Matching Cohen's *d*	After Matching Cohen's *d*
		CBT	PBT	CBT	PBT		
School	*N*	4388	1869	2227	1443	-	-
% African American	Mean	16.97	32.05	17.45	26.26	0.414	0.248
% Hispanic	Mean	21.99	20.85	18.63	22.26	0.035	0.111
% White	Mean	51.60	39.08	57.25	43.62	0.306	0.330
% Female	Mean	48.67	49.17	47.87	49.52	0.015	0.050
% EDS	Mean	44.30	57.41	42.79	56.18	0.321	0.327
% EL	Mean	11.04	14.42	11.35	15.00	0.133	0.140
% SWD	Mean	8.43	7.81	5.62	5.28	0.031	0.019
Students	*N*	11972	10473	5807	7981	-	-
						0.179*	0.175*

Note: *Average Cohen's *d* of all covariates

Table 3.9 Characteristics of Schools before and after Propensity Score Matching for ELA/L Grade 3 Form Pair with State A Excluded

Variable	Statistics	Overall					
		Before Matching		After Matching		Before Matching Cohen's *d*	After Matching Cohen's *d*
		CBT	PBT	CBT	PBT		
School	*N*	3399	940	1243	627	-	-
% African American	Mean	18.95	31.84	23.22	26.88	0.352	0.098
% Hispanic	Mean	22.37	15.60	17.00	14.94	0.221	0.071
% White	Mean	48.87	44.08	54.32	50.77	0.118	0.085
% Female	Mean	48.83	49.30	47.58	49.00	0.015	0.041
% EDS	Mean	42.68	44.20	37.13	40.76	0.037	0.089
% EL	Mean	9.96	9.76	8.70	8.66	0.009	0.002
% SWD	Mean	10.87	15.37	10.04	12.10	0.187	0.088
Students	*N*	9248	5268	3104	3366	-	-
						0.134*	0.068*

Note: *Average Cohen's *d* of all covariates

Table 3.10 Characteristics of Schools before and after Propensity Score Matching for ELA/L Grade 7 Form Pair

Variable	Statistics	Overall					
		Before Matching		After Matching		Before Matching Cohen's *d*	After Matching Cohen's *d*
		CBT	PBT	CBT	PBT		
School	*N*	3521	1045	1396	702	-	-
% African American	Mean	19.72	28.63	21.15	27.28	0.244	0.160
% Hispanic	Mean	19.37	13.37	11.78	11.51	0.205	0.010
% White	Mean	51.45	41.64	57.69	43.83	0.241	0.328
% Female	Mean	49.43	43.06	47.11	44.69	0.207	0.074
% EDS	Mean	46.98	51.83	49.71	50.96	0.121	0.031
% EL	Mean	4.59	5.77	2.03	3.68	0.073	0.137
% SWD	Mean	8.12	22.86	6.32	12.75	0.558	0.287
Students	*N*	14990	10854	5233	6563	-	-
						0.236*	0.147*

Note: *Average Cohen's *d* of all covariates

Table 3.11 Characteristics of Schools before and after Propensity Score Matching for ELA/L Grade 9 Form Pair

Variable	Statistics	Overall					
		Before Matching		After Matching		Before Matching Cohen's *d*	After Matching Cohen's *d*
		CBT	PBT	CBT	PBT		
School	*N*	1286	402	495	275	-	-
% African American	Mean	15.94	17.60	16.25	16.15	0.053	0.003
% Hispanic	Mean	15.25	5.74	4.17	4.38	0.414	0.014
% White	Mean	57.76	47.52	63.09	52.81	0.248	0.246
% Female	Mean	49.41	48.07	47.98	47.75	0.042	0.007
% EDS	Mean	44.69	39.40	38.35	37.14	0.133	0.030
% EL	Mean	3.07	1.60	0.81	0.92	0.146	0.018
% SWD	Mean	9.28	15.86	7.85	13.77	0.264	0.248
Students	*N*	4615	4512	1548	2813	-	-
						0.186*	0.081*

Note: *Average Cohen's *d* of all covariates

Table 3.12 Characteristics of Schools before and after Propensity Score Matching for Mathematics Grade 5 Form Pair for State A.

Variable	Statistics	State A					
		Before Matching		After Matching		Before Matching Cohen's *d*	After Matching Cohen's *d*
		CBT	PBT	CBT	PBT		
School	*N*	1012	708	978	590	-	-
% African American	Mean	10.89	36.32	10.98	31.03	0.696	0.561
% Hispanic	Mean	19.76	27.61	19.58	26.84	0.220	0.205
% White	Mean	59.69	29.83	60.81	35.02	0.739	0.630
% Female	Mean	52.18	50.56	51.71	51.99	0.048	0.008
% EDS	Mean	47.29	70.86	47.68	66.49	0.597	0.469
% EL	Mean	2.30	4.66	2.03	3.57	0.198	0.148
% SWD	Mean	0.10	0.11	0.10	0.13	0.003	0.010
Students	*N*	2475	3676	2384	2958	-	-
						0.357*	0.290*

Note: *Average Cohen's *d* of all covariates

Table 3.13 Characteristics of Schools before and after Propensity Score Matching for ELA/L 3 Form Pair for State A

Variable	Statistics	State A					
		Before Matching		After Matching		Before Matching Cohen's *d*	After Matching Cohen's *d*
		CBT	PBT	CBT	PBT		
School	*N*	989	929	984	816	-	-
% African American	Mean	10.17	32.26	10.17	25.78	0.649	0.486
% Hispanic	Mean	20.69	26.17	20.69	27.88	0.159	0.207
% White	Mean	60.95	34.02	60.95	38.12	0.673	0.567
% Female	Mean	48.10	49.02	48.23	49.92	0.029	0.052
% EDS	Mean	49.83	70.78	49.93	68.03	0.547	0.468
% EL	Mean	14.75	19.14	14.69	19.88	0.154	0.182
% SWD	Mean	0.05	0.16	0.05	0.05	0.060	0.002
Students	*N*	2724	5205	2703	4615	-	-
						0.324*	0.281*

Note: *Average Cohen's *d* of all covariates

3.5 Summary

In order to select a sufficiently large sample for statistical analyses, a less stringent criteria of Cohen's *d* less than 0.2 was used to evaluate matched samples. For six out of the selected eight core form pairs, matched samples with the balance of the covariate distributions between the CBT and PBT groups met the less than 0.2 Cohen's *d* criteria, including all available states. For the Mathematics grade 5 and ELA/L grade 3 core form pairs, matched samples between the CBT and PBT groups met the Cohen's d criteria, excluding State A schools. To evaluate analyses results in matched samples with or without State A schools, all statistical and psychometric analyses appearing in the subsequent chapters were conducted using both samples. Because no meaningful differences in the results were seen for the two samples with or without State A schools, in the subsequent chapters only the results based on the matched samples that did not exclude State A students were reported. In summary, given that matching was conducted on demographic data that were less predictive than prior achievement and that a more stringent criterion could not be applied without it severely impacting the overall sample size that was required to perform the statistical and psychometric analyses, the matching results were less than ideal and would have impacted the analyses conducted on the matched samples. Therefore, results from the current mode comparability study are not conclusive; the results should be considered as preliminary and descriptive.

Section 4: Analyses and Results Pertaining to Construct Invariance

4.1 Overview

Analyses in this section were designed to assess the degree to which the same construct is measured by the CBT and PBT versions of the 2015 PARCC operational assessments. These analyses focus on the internal structure of each test and the degree to which the structures are similar. As noted in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014, p. 16), "Analysis of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based."

4.2 Effect Size Analyses

4.2.1 Objectives

The goal of the analyses appearing in this section was to evaluate the similarity of common item performance across modes in selected grade levels and form pairs using classical test theory methods. It is important to know whether an item's relative difficulty is consistent across modes. Lack of consistency across modes might indicate that an item is measuring a different attribute and might warrant further study of its qualitative characteristics.

4.2.2 Method

Summary statistics obtained for the common items administered in each mode were calculated. For each item, the average item score was calculated as the average number of points earned; the average number of points was then rescaled by dividing its value by the maximum score points available so that the difficulty interpretation would be consistent for both dichotomous and polytomous items. This is ordinarily referred to as a "p-value" for a dichotomous item, and this term is being extended to polytomous items in this report.

To examine the consistency of the items' relative difficulties across the CBT and PBT modes, the *z*-scores of *p-values* from each mode were correlated and the effect sizes of *p-value* differences between the test modes were calculated. *Z*-scores were calculated using the following formula,

$$z_{im} = \frac{p_{im} - \overline{p}_m}{s_{pm}} \tag{2}$$

where, p_{im} is the *p-value* for item *i* within a test condition *m (m=paper or computer for mode comparability)* $\overline{p}_m$ is the mean of the items in test condition *m*, and s_{pm} is the standard deviation of the *p-values* of the items in test condition *m*.

Effect sizes were next calculated using the following formula.

$$Effect\ Size = \frac{p_{CBT} - p_{PBT}}{\sqrt{(s_{CBT}^2 + s_{PBT}^2)/2}} \tag{3}$$

where p_{PBT} and p_{CBT} are the *p-values* from CBT and PBT modes respectively*,* and s_{CBT} and s_{PBT} are the standard deviations of the *p-values* of the items in different test modes. Items were classified as outliers if an effect size of *p-value* difference was greater than the absolute value of 0.20, as defined by Cohen (1988). Items classified as "outliers" were considered less comparable across modes and required further examination by content experts. The characteristics of these items were summarized to investigate whether certain item types function more differently cross modes than other item types. Items classified as outliers were not removed from IRT calibrations or equating.

4.2.3 Results

Tables 4.1 and 4.4 provide the mean and median *p-values* for the common items appearing in both modes, summarized for each subject and grade level evaluated in this study. ELA/L common items were slightly easier on PBT for grades 3 and 9 but easier on computer for grade 7, with the median *p-value* differences (CBT - PBT) ranging from -0.03 to 0.02. Even though the overall median *p-value* differences were small, large differences were found for the prose constructed response (PCR) common items. The average *p-value* differences of PCR trait items ranged from -0.13 to -0.03, all favoring the PBT mode. In the 2015 PARCC spring administration there were three type of PCR trait items, Reading Comprehension (RD), Writing Expression (WE), Writing Knowledge and Conventions (WKL). The largest differences occurred for grades 3 and 9 WKL items. The median *p-value* differences among non-PCR items were, on the other hand, much smaller ranging from 0.01 to 0.02. Refer to Table 4.2 for *p-value* summary statistics for ELA/L PCR trait items and Table 4.3 for *p-value* summary statistics for ELA/L non-PCR items. Common item performance was better on computer than on paper for mathematics grades 5, 7 and Algebra I with median *p-value* differences (CBT - PBT) of 0.04, 0.09 and 0.02, respectively. However, for the high school Geometry and Algebra II tests, common items were easier on paper than on computer with median *p*-value differences (CBT - PBT) of -0.08 and -0.03, respectively.

Table 4.1 Average *p-Values* across Administration Modes for ELA/L Common Items

Grade	Mode	N	Mean	Min	Max	SD	Median
3	CBT	31	0.42	0.14	0.70	0.17	0.40
	PBT	31	0.45	0.21	0.73	0.15	0.43
7	CBT	47	0.44	0.23	0.77	0.15	0.42
	PBT	47	0.42	0.21	0.75	0.13	0.40
9	CBT	44	0.47	0.11	0.84	0.18	0.46
	PBT	44	0.51	0.13	0.88	0.18	0.49

Table 4.2 Average *p-Value*s across Administration Modes for ELA/L PCR Common Items

Grade		Mode	N	Mean	Min	Max	SD
3	RD	CBT	2	0.19	0.17	0.21	0.03
		PBT	2	0.27	0.24	0.29	0.03
	WE	CBT	3	0.20	0.14	0.30	0.09
		PBT	3	0.28	0.21	0.38	0.09
	WKL	CBT	3	0.29	0.22	0.36	0.07
		PBT	3	0.37	0.31	0.43	0.06
7	RD	CBT	2	0.27	0.23	0.31	0.06
		PBT	2	0.31	0.26	0.36	0.07
	WE	CBT	3	0.25	0.23	0.30	0.04
		PBT	3	0.28	0.24	0.35	0.06
	WKL	CBT	3	0.35	0.32	0.41	0.05
		PBT	3	0.39	0.35	0.46	0.06
9	RD	CBT	2	0.29	0.24	0.34	0.07
		PBT	2	0.39	0.35	0.42	0.05
	WE	CBT	3	0.30	0.23	0.34	0.06
		PBT	3	0.41	0.35	0.46	0.06
	WKL	CBT	3	0.39	0.33	0.43	0.05
		PBT	3	0.52	0.47	0.56	0.05

Table 4.3 Average *p-Values* across Administration Modes for ELA/L non-PCR Common Items

Grade	Mode	N	Mean	Min	Max	SD	Median
3	CBT	23	0.48	0.22	0.70	0.14	0.49
	PBT	23	0.49	0.22	0.73	0.14	0.47
7	CBT	39	0.47	0.23	0.77	0.15	0.45
	PBT	39	0.43	0.21	0.75	0.14	0.43
9	CBT	36	0.51	0.11	0.84	0.19	0.52
	PBT	36	0.52	0.13	0.88	0.19	0.51

Table 4.4 Average *p-Values* across Administration Modes for Mathematics Common Items

Subject/Grade	Mode	N	Mean	Min	Max	SD	Median
5	CBT	23	0.51	0.04	0.85	0.25	0.59
	PBT	23	0.50	0.03	0.84	0.24	0.55
7	CBT	21	0.36	0.02	0.81	0.28	0.28
	PBT	21	0.30	0.01	0.73	0.25	0.19
Algebra I	CBT	20	0.22	0.02	0.70	0.22	0.10
	PBT	20	0.21	0.02	0.69	0.22	0.08
Geometry	CBT	30	0.29	0.02	0.76	0.22	0.22
	PBT	30	0.34	0.01	0.77	0.23	0.30
Algebra II	CBT	19	0.25	0.00	0.66	0.22	0.20
	PBT	19	0.29	0.00	0.72	0.23	0.23

As shown in Tables 4.5 and 4.6, the *z*-scores correlations between CBT and PBT modes were consistently high, nearing 1.0, except for Geometry ($r = 0.61$). After examining the item statistics, a multiple choice item in Geometry was identified that was answered correctly by 31% of the students taking the test online but less than 2% of the students taking the test on paper. After removing this outlier item, the correlation of item *z*-scores improved to 0.96. Assessment development experts reviewed the item but did not find any issues that might explain the performance difference across modes.

Items with effect sizes greater than 0.2 were also flagged. Only one item was flagged for ELA/L grade 7 whereas a much larger percentage of items were flagged for ELA/L grades 3 and 9, 29% and 18% respectively. Detailed information about characteristics of items flagged for effect sizes greater than 0.2 is included in Appendix A and descriptions of item types are listed in Appendix D. Table A.1 indicates that, for ELA/L, the majority of the items flagged for large effect size had low cognitive complexity, had point values of 3 or 4, were an Extended Text Interaction item types, were 3- or 4-part multiple choice single select items, or were Literary Analysis Task types. All flagged items were reviewed by assessment development content experts and were found to be acceptable from a content perspective.

Fewer items were flagged for mathematics grade 5 and none for Algebra I, but higher percentages of items were flagged for mathematics grade 7, Geometry, and Algebra II. Over 40% of Geometry common items were flagged for large effect sizes. The item with the largest effect size (0.86) across all subjects/grades was also from a Geometry test. This was the same item that caused the low correlation between item *z*-scores. Table A.2 suggests that both one- and two-part constructed response items were more likely to perform differently across modes, as well as other interaction item types, items with higher cognitive complexity, judgment based response types, Type 1 - 4 point items, or Type 3 - 3 and 6 point items. Content experts reviewed the flagged items and found that they tended to be items that required students show their work, provide justification for their response, or draw on a graph to solve the problem.

Table 4.5 Common Items across Modes in ELA/L: Z-Score Correlations and Items Flagged for Large *P-value* Differences

Grade	Number of Items	Number of Items Flagged	Percentage Flagged	Largest effect Size Difference	*Z*-score Correlation
3	31	9	29%	-0.37	0.98
7	47	1	2%	0.27	0.99
9	44	8	18%	-0.52	0.98

Table 4.6 Common Items across Modes in Mathematics: *Z*-Score Correlations and Items Flagged for Large *P-value* Differences

Grade	Number of Items	Number of Items Flagged	Percentage Flagged	Largest effect Size Difference	*Z*-score Correlation
5	23	1	4%	-0.49	0.99
7	21	5	24%	0.29	0.99
Algebra I	20	0	0%	0.12	0.98
Geometry	30	13	43%	0.86	0.61
Algebra II	19	6	32%	-0.44	0.99

4.2.4 Summary

Even though the median differences in difficulty tended to be small for ELA/L tests, the differences that were sizable for the PCR trait items all favored PBT students. The numbers of items flagged for *p-value* effect sizes larger than 0.2 also varied across grades and subjects. Geometry had the largest percentage of items flagged for effect size. The *z*-score correlations were generally high.

4.3 Differential Item Functioning

4.3.1 Objectives

As defined in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014, p.51), "*Differential item functioning* (DIF) is said to occur when equally able test takers differ in their probabilities of answering a test item correctly as a function of group membership." Although significant research has been devoted to supporting development of innovative items and item types, evidence should be collected to evaluate whether the skills measured by these items are the same across the CBT and PBT test modes. The goal of this section is to evaluate the degree in which mode of administration might introduce construct irrelevant variance at the item level for test takers of equal ability. The results of the DIF analyses will further broaden the understanding of the performance of PARCC items as well as inform future test development efforts.

4.3.2 Method

The analysis of mode DIF was carried out using the Mantel-Haenszel DIF procedure (MH DIF; Dorans & Holland, 1993; Mantel & Haenszel, 1959) for selected response (SR) items, and a combination of the Mantel-Haenszel (MH) method for ordinal variables and standardization procedures (Dorans & Schmitt,

1991) for constructed response (CR) items. For the standardization procedure, the DIF statistic was based on the standardized mean difference (SMD) in average item scores between members of two groups (i.e., modes) who have been matched on their ability. All DIF analyses conducted in the study used the theta estimates of students obtained from CBT and PBT scaling of 2015 PARCC operational test as the matching criteria.

Based on the DIF statistics and significance tests, items were classified into one of three categories: Category A items contain negligible DIF, Category B items exhibit slight to moderate DIF, and Category C items have moderate to large values of DIF. Negative values imply that, conditional on the matching variable, the focal group (PBT) has a lower mean item score than the reference group (CBT). In contrast, a positive value implies that, conditional on total common item score; the reference group (CBT) has lower mean item score than the focal group (PBT). Current practice only considers Category C DIF to be a potential threat to item fairness and to warrant further investigation (Educational Testing Service, 2002). Tables 4.7 and 4.8 provide the flagging criteria for SR and CR items, respectively.

Table 4.7 DIF Categories for Selected-Response Items

DIF Category	Criteria
A (negligible)	Absolute value of the MH D-DIF is not significantly different from zero, or is less than one.
B (slight to moderate)	1. Absolute value of the MH D-DIF is significantly different from zero but not from one, and is at least one; OR 2. Absolute value of the MH D-DIF is significantly different from one, but is less than 1.5. Positive values are classified as “B+” and negative values as “B-”.
C (moderate to large)	Absolute value of the MH D-DIF is significantly different from one, and is at least 1.5. Positive values are classified as “C+” and negative values as “C-.”

Table 4.8 DIF Categories for Polytomously Scored Items

DIF Category	Criteria
A (negligible)	Mantel Chi-square *p*-value > 0.05 and $|SMD/SD| \leq 0.17$
B (slight to moderate)	Mantel Chi-square *p*-value < 0.05 and $|SMD/SD| > 0.17$
C (moderate to large)	Mantel Chi-square *p*-value < 0.05 and $|SMD/SD| > 0.25$

4.3.3 Results

Table 4.9 summarizes the DIF results across all grade levels and subjects. Appendix B provides grade- and test-specific DIF results for ELA/L and mathematics assessments. The characteristics of items that were flagged for C-level DIF are listed in Appendix C and descriptions of PARCC item types are listed in Appendix D. Overall, across grade levels, there were very few instances of C-level DIF based on the Mantel-Haenszel and SMD approaches with flagging rates ranging from 0% to 15% of the items within each grade/test. The larger percentage of mode DIF occurred for mathematics, especially for high school mathematics assessments. Generally, there was a slightly larger percentage of items favoring the CBT mode (C-) over PBT across mathematics grades 3 to 8. Nevertheless the directions of C-level DIF items were inconsistent across different subjects/grade levels of high school mathematics tests. The characteristics of mathematics items flagged for C-DIF shown Table C.2 suggest that “Fill-in-the-Blank” and multiple choice multiple select items were more likely to function differently across modes. Moreover, Text Entry Interaction types and judgment response types were more likely to be flagged for C-Level DIF. Content experts reviewed the items flagged for C-DIF and found that these items tended to require students to show their work, provide justification for their response, or draw on a graph to solve the problem. In addition the CBT “Fill-in-the-Blank” items did not allow students to use as many characters as students testing on paper; this may have impacted students’ performance on these item types. Only four common items in ELA/L assessments were identified as having C-level DIF and the specific features were not described given the small number.

Table 4.9 Summary of Mantel-Haenszel/SMD DIF Results across Assessments

Test	DIF Category	Mantel-Haenszel/SMD	
		Total Number of Common Items	Percentage
ELA/L	A	576	94%
	B-[a]	6	1%
	B+[b]	24	4%
	C-[a]	2	<1%
	C+[b]	2	<1%
	Total	610	
Mathematics	A	819	86%
	B-[a]	40	4%
	B+[b]	39	4%
	C-[a]	29	3%
	C+[b]	23	2%
	Total	950	

Note: [a]"B-" and "C-" DIF indicates the item favors the CBT group.
[b]"B+" and "C+" indicates the item favors the PBT group.

4.3.4 Summary

The results from the DIF analyses indicated that the level of construct irrelevant variance introduced by test mode varied by content area. There were a small percentage of items that were flagged for performing differently across modes conditional on ability for ELA/L. There was a larger percentage of mathematics items flagged for DIF, particularly for high school assessments. Consistent with results from the field test mode comparability study, many items flagged for C-level DIF in grades 3 to 8 mathematics assessments were "fill-in-the-blank" items. For high school mathematics tests, most of the C-level DIF items were constructed response items and "fill-in-the-blank" items. Compared to the DIF results obtained from the 2014 field test mode comparability study, the current study had a smaller percentages of items flagged as C-DIF for ELA/L (0.7% vs. 1.6%) and mathematics (5% vs. 9%). For ELA/L most of the C-level DIF items in the field test were either not included in 2015 operational tests or were included in the operation tests but not used as common items across modes. In addition, several items appearing in the field tests were classified as C-DIF because responses were missing from one mode. This issue was not found in the 2015 operational tests. For mathematics, compared to the 2014 field test, the 2015 operational tests had a smaller percentage of "fill-in-the-blank" items which are in general more likely to be flagged for C-DIF. Moreover, in the process of transitioning from the field test phase to operational tests, much was learned about the items and assessments which helped improved the overall quality of the operational assessments.

Section 5: Analyses and Results Pertaining to IRT Item Parameter Estimates

5.1 Overview

Section 5 describes the item response theory (IRT) analyses that were used to evaluate both the comparability of parameter estimates across testing modes and the sensitivity of these parameter estimates across various calibration approaches. The data used in this study are from propensity score matching of CBT and PBT students taking selected form pairs. Refer to Section 2.2 for information about form pairs and grade levels selected for the mode comparability study. Note that the study only considers the operational items within core operational test forms. The field test items on each form were not included in the study.

There were two high-level steps in the IRT analyses for this study. First, selected CBT and PBT full summative forms (Refer to Table 2.2) were calibrated separately. Second, item parameters for common items between CBT and PBT forms were estimated in the presence of and in the absence of uncommon items (i.e., unique items in CBT and PBT forms). Details are presented below.

5.2 Data Manipulation

Prior to performing the calibration, the following data treatment occurred. Based on classical item analysis (IA) results from aggregated data across forms, items were excluded from the IRT calibration using the following criteria:

1. Exclude items with a weighted polyserial correlation less than 0.0.
2. Exclude items with an average item score of 0.0.
3. Exclude items where 100% of the students have the same item score, such as:
    a. 100% omitted the item,
    b. 100% received the same score,
    c. 100% of the responses were at the same score after collapsing score categories due to low frequencies, or
    d. 100% of the responses were not presented or not reached.
4. Exclude items with insufficient sample sizes for the selected IRT model combinations (i.e., 500 for the 2PL/GPC).
5. Exclude items with high omit rates (i.e., greater than 50%) on one or more forms.

Additionally, if an item prevented the software from converging, this was addressed by either collapsing that item's score categories or exclude the item from calibration. An example of a collapsed item might involve reducing it from 5 points: 0, 1, 2, 3, 4, to 4 points: 0, 1, 2, 3. The item exclusion criteria were consistent with the 2015 PARCC operational test calibration (PARCC, 2016). No items were dropped from calibration due to difficulties with convergence in this study but several items' score categories were collapsed to improve estimation (Refer to Section 5.4.3).

5.3 Calibration of PCR Traits

Prose Constructed Response (PCR) trait items in ELA/L assessments were calibrated at the trait score level rather than the aggregated total score level. There are three PCR trait items in each core operational test form. Given the smaller sample size of matched CBT and PBT students in the mode comparability study (e.g., less than 2,000 for some of the subjects/grade levels), each PCR trait was calibrated jointly with all non-PCR items but separately from the other two PCR traits using the full-sample data. After the three PCR trait calibrations, parameter estimates from the Writing Expression and Writing Knowledge and Conventions trait calibrations were linked to the scale of the Reading Comprehension trait using the non-PCR items as linking items using the Stocking and Lord approach. The STUIRT (Kim & Kolen, 2004) software was used in the scaling process. All non-PCR items were included as linking items for the PCR trait scaling.

5.4 Separate Calibration of Common Items

5.4.1 Objectives

The goal of the analysis described in this section was to evaluate whether the estimated difficulty and discrimination parameters based on separate within-mode IRT calibrations indicate that the items measure the same construct. Results from this analysis have implications for item banking as well as for placing item parameter estimates derived from separate calibrations onto a common scale. Specifically, if common item parameter estimates based on separate within-mode calibrations can be assumed, then only one set of item parameters will be needed. In terms of scaling, since there may be items that function differently across modes or mode–specific items (e.g., technology enhanced items only administered online), there is a need to place the item parameter estimates from such items onto the common scale established by calibration of CBT forms. High correlations between items common across modes will support linking. For the 2-parameter logistic (2PL) model, the minimally recommended correlations are .85 and .95 for the discrimination and difficulty parameters, respectively (H. Huynh, personal communication, May 18, 2015).

5.4.2 Method

The selected CBT and PBT full summative forms were independently calibrated. The difficulty and discrimination parameter estimates of common items between CBT and PBT forms from independent calibrations were correlated and plotted. The analysis was based on using the 2-parameter logistic/generalized partial credit model (2PL/GPC) combination which is consistent with the IRT model used for 2015 PARCC operational tests. The analysis was performed using commercial PARSCALE (Muraki & Bock, 2003).

5.4.3 Results

There were three cases in which categories were collapsed for common items in the PBT data due to the small sample sizes for the highest category. The categories for those common items were also collapsed for the CBT data to support calibration. For Algebra 2, one item’s categories 4 and 5 and another item’s categories 6 and 7 were collapsed. For Geometry, one item’s categories 3 and 4 were collapsed. Two PCR traits (one Reading Comprehension trait and one Writing Knowledge and Conventions trait) in ELA/L grade 3 and one Comprehension trait in ELA/L grade 7 were calibrated with the highest category collapsed with the next lower level category. Table 5.1 provides the correlation summaries across grades for difficulty and discrimination parameter estimates. Figures 5.1 – 5.3 provide the corresponding correlation plots for difficulty and discrimination parameter estimates for ELA/L and Figures 5.4 – 5.8 illustrate the corresponding correlations for difficulty and discrimination parameter estimates for mathematics. For both ELA/L and mathematics, the correlations associated with the IRT difficulty parameter estimates, as is typically the case with IRT estimation, tended to be stronger than those for the discrimination parameter estimates. The identity line in the graphs help identify the differences between CBT and PBT parameter estimates. For most of the grade levels Item discriminations tended to be measured similarly across mode whereas there were differences for difficulties for the majority of grade levels, especially for ELA/L grade 3 and grade 9, mathematics grade 7, Algebra II and Geometry. Most items' b parameters were higher (more difficult) based on the CBT calibrations than those based on the PBT calibrations for ELA/L grade 3 and grade 9 and Geometry whereas the opposite was true for mathematics grade 7, which was consistent with classical item analyses results as described in Tables 4.1 and 4.4.

Table 5.1 Correlations between Modes of Discrimination and Difficulty Parameter Estimates for Common Items

	Grade	Number of Items	Discrimination	Difficulty
ELA/L	3	31	.94	.94
	7	47	.96	.97
	9	44	.95	.98
	Grade/Subject	Number of Items	Discrimination	Difficulty
Mathematics	5	23	.94	.98
	7	21	.96	.95
	Algebra I	20	.92	.96
	Geometry	30	.92	.72*
	Algebra II	19	.90	.96

Note: *Refer to discussion of this correlation in the Section 5.4.4.

5.4.4 Summary

For most grade levels and content domains, the estimated difficulties and discriminations, based on separate within mode calibrations of the items, tended to have a high degree of coherence as evidenced by the high correlations. Based on the recommended correlation thresholds for linking items from separate calibrations onto a common scale, linking PBT items onto the same scale as CBT items should not present problems at most grade levels. For grades and subjects, where the correlation thresholds are not met, additional refinements to the common item set, such as removing outlier items, could be performed to support linking. In the case of Geometry, for example, removing the problematic item that performed dramatically different across modes (Refer to Section 4.2.3 for details) changed the correlation of the difficulty parameter estimates from .72 (below the recommend correlation threshold of .95) to .96 (above the threshold). Differences in IRT difficulties were found for the majority of grade levels, especially for ELA/L grade 3 and grade 9, mathematics grade 7, Algebra II and Geometry, which were consistent with item *p-value* differences in classical item analyses results

5.5 Joint Calibration of Common Items

5.5.1 Objectives

The goal of this analysis was to evaluate the sensitivity of calibration results in the presence and absence of the non-common items (i.e., items unique to a particular mode).

5.5.2 Method

There were three possible item groups across the modes:

CM: Common items

C1: Condition 1-specific items (e.g., CBT items)

C2: Condition 2-specific items (e.g., PBT items)

Based on the item groupings, there were four conditions considered in which data were pooled from both modes for calibration in order to estimate parameters for CM:

(1) Calibrate CM items only
(2) Joint calibration of CM+ C1 items
(3) Joint calibration of CM+ C2 items
(4) Joint calibration of CM+ C1+C2 items

The item parameter estimates corresponding to common items were correlated across the four conditions. Strong correlations among the estimates for these conditions would provide evidence that the same construct is being measured. Additionally, it was important to evaluate whether calibrating items common to both modes would be impacted if they were also calibrated with items unique to each mode. Refer to Section 5.2.2 for the description of the data treatment and software package used to conduct the analysis.

5.5.3 Results

Tables 5.2 and 5.3 present the correlations of parameter estimates for ELA/L and mathematics, respectively. Overall, the difficulty and discrimination estimates were minimally impacted by the different approaches used to calibrate items. The correlations almost always ranged from .99 to 1.00. As was the case with the separate calibrations, for both ELA/L and mathematics, the correlations associated with the IRT difficulty parameter estimates appeared to be stronger than those for the discrimination parameter estimates. There was a slight degradation in the correlations when all items were calibrated together for both ELA/L and mathematics.

5.5.4 Summary

Overall, the four calibration approaches evaluated yielded very small differences in the results. Moreover, there was not enough evidence to indicate the differences in the calibration results would have a significant impact on psychometric analyses.

Table 5.2 Impact of Calibration Conditions for Difficulty and Discrimination Parameter Estimates for ELA/L

		Discrimination				Difficulty			
Test	Calibration Condition	CM Only	CM + CBT	CM + PBT	CM+CBT +PBT	CM Only	CM + CBT	CM + PBT	CM+CBT +PBT
Grade 3	CM Only	1	.	.	.	1	.	.	.
	CM + CBT	.99982	1	.	.	.99999	1	.	.
	CM + PBT	.99935	.99955	1	.	.99999	.99997	1	.
	CM + CBT + PBT	.99890	.99942	.99986	1	.99998	.99999	.99999	1
Grade 7	CM Only	1	.	.	.	1	.	.	.
	CM + CBT	.99997	1	.	.	.99999	1	.	.
	CM + PBT	1	.99998	1	.	1	.99999	1	.
	CM + CBT + PBT	.99991	.99998	.99997	1	.99999	.99999	.99999	1
Grade 9	CM Only	1	.	.	.	1	.	.	.
	CM + CBT	.99989	1	.	.	.99998	1	.	.
	CM + PBT	.99955	.99974	1	.	.99999	.99998	1	.
	CM + CBT + PBT	.99919	.99958	.99991	1	.99995	.99999	.99998	1

Note: Correlations are displayed in five digits in this table to show that they were not all equal to 1.00.

Table 5.3 Impact of Calibration Conditions for Difficulty and Discrimination Parameter Estimates for Mathematics

		Discrimination				Difficulty			
Test	Calibration Condition	CM Only	CM + CBT	CM + PBT	CM + CBT + PBT	CM Only	CM + CBT	CM + PBT	CM + CBT + PBT
Grade 5	CM Only	1	.	.	.	1	.	.	.
	CM + CBT	.99484	1	.	.	.99985	1	.	.
	CM + PBT	.99763	.99539	1	.	.99994	.99980	1	.
	CM + CBT + PBT	.99241	.99808	.99693	1	.99982	.99994	.99988	1
Grade 7	CM Only	1	.	.	.	1	.	.	.
	CM + CBT	.99929	1	.	.	.99991	1	.	.
	CM + PBT	.99687	.99688	1	.	.99950	.99949	1	.
	CM + CBT + PBT	.99634	.99750	.99950	1	.99943	.99957	.99993	1
Algebra I	CM Only	1	.	.	.	1	.	.	.
	CM + CBT	.99933	1	.	.	.99979	1	.	.
	CM + PBT	.99606	.99621	1	.	.99964	.99941	1	.
	CM + CBT + PBT	.99535	.99662	.99951	1	.99951	.99967	.99981	1
Geometry	CM Only	1	.	.	.	1	.	.	.
	CM + CBT	.99913	1	.	.	.99995	1	.	.
	CM + PBT	.99944	.99893	1	.	.99997	.99996	1	.
	CM + CBT + PBT	.99833	.99948	.99920	1	.99989	.99998	.99995	1
Algebra II	CM Only	1	.	.	.	1	.	.	.
	CM + CBT	.99925	1	.	.	.99988	1	.	.
	CM + PBT	.99769	.99605	1	.	.99963	.99950	1	.
	CM + CBT + PBT	.99835	.99818	.99924	1	.99961	.99970	.99989	1

Note. Correlations are displayed in five digits in this table to show that they were not all equal to 1.00.

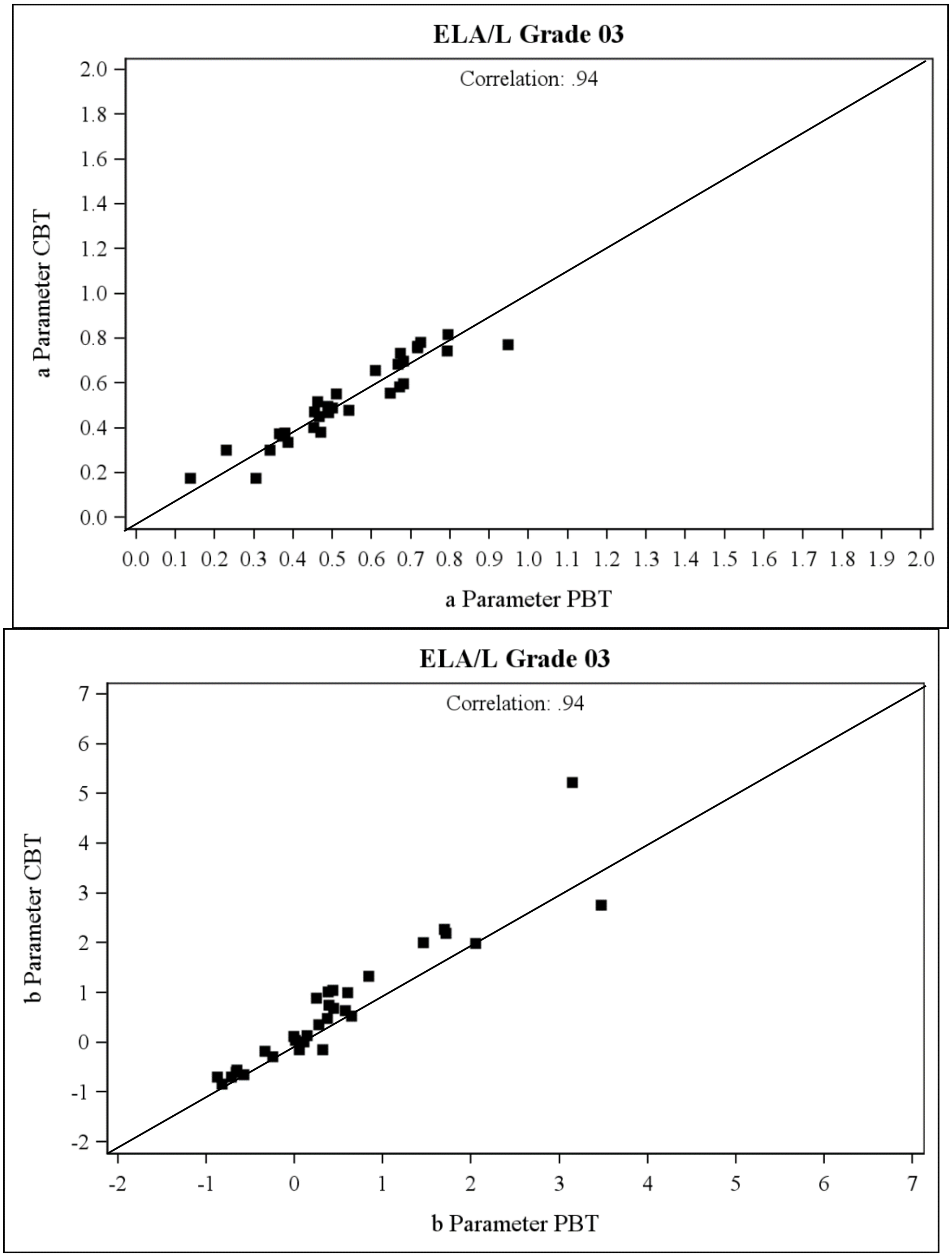


Figure 5.1. Correlation between Difficulty and Discrimination Parameter Estimates across Modes for ELA/L Grade 3.

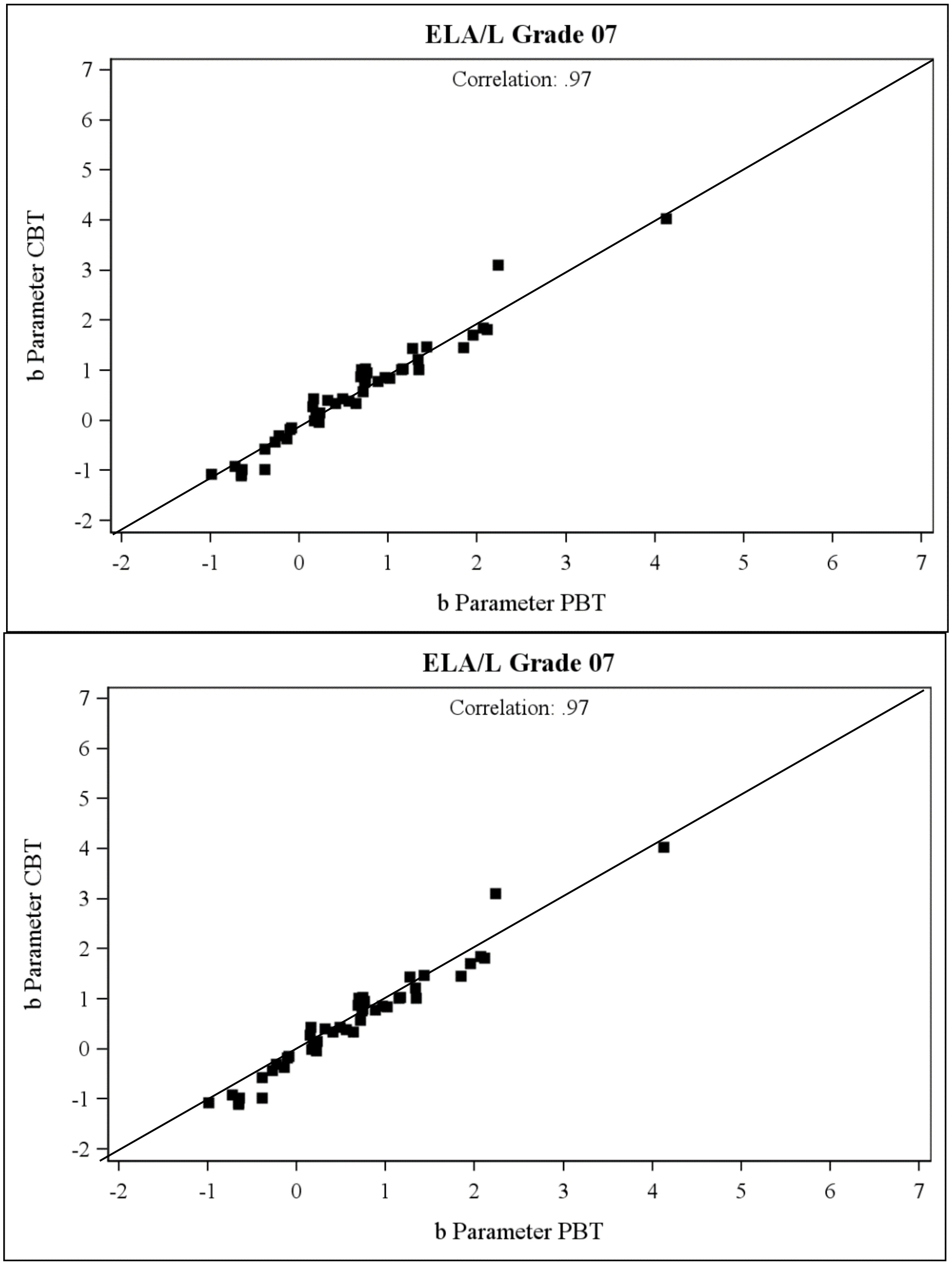


Figure 5. 2. Correlation between Difficulty and Discrimination Parameter Estimates across Modes for ELA/L Grade 7.

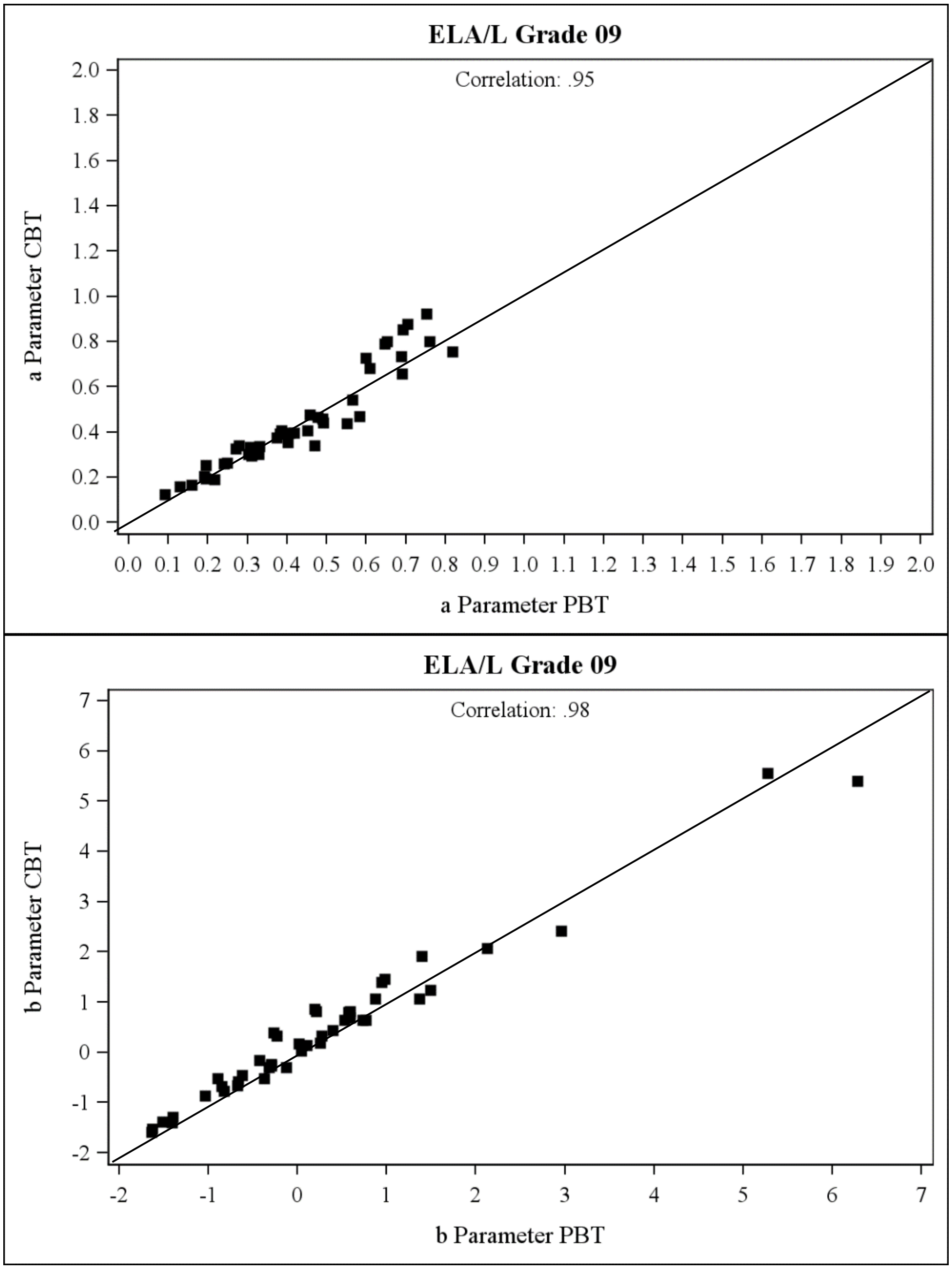


Figure 5.3. Correlation between Difficulty and Discrimination Parameter Estimates across Modes for ELA/L Grade 9.

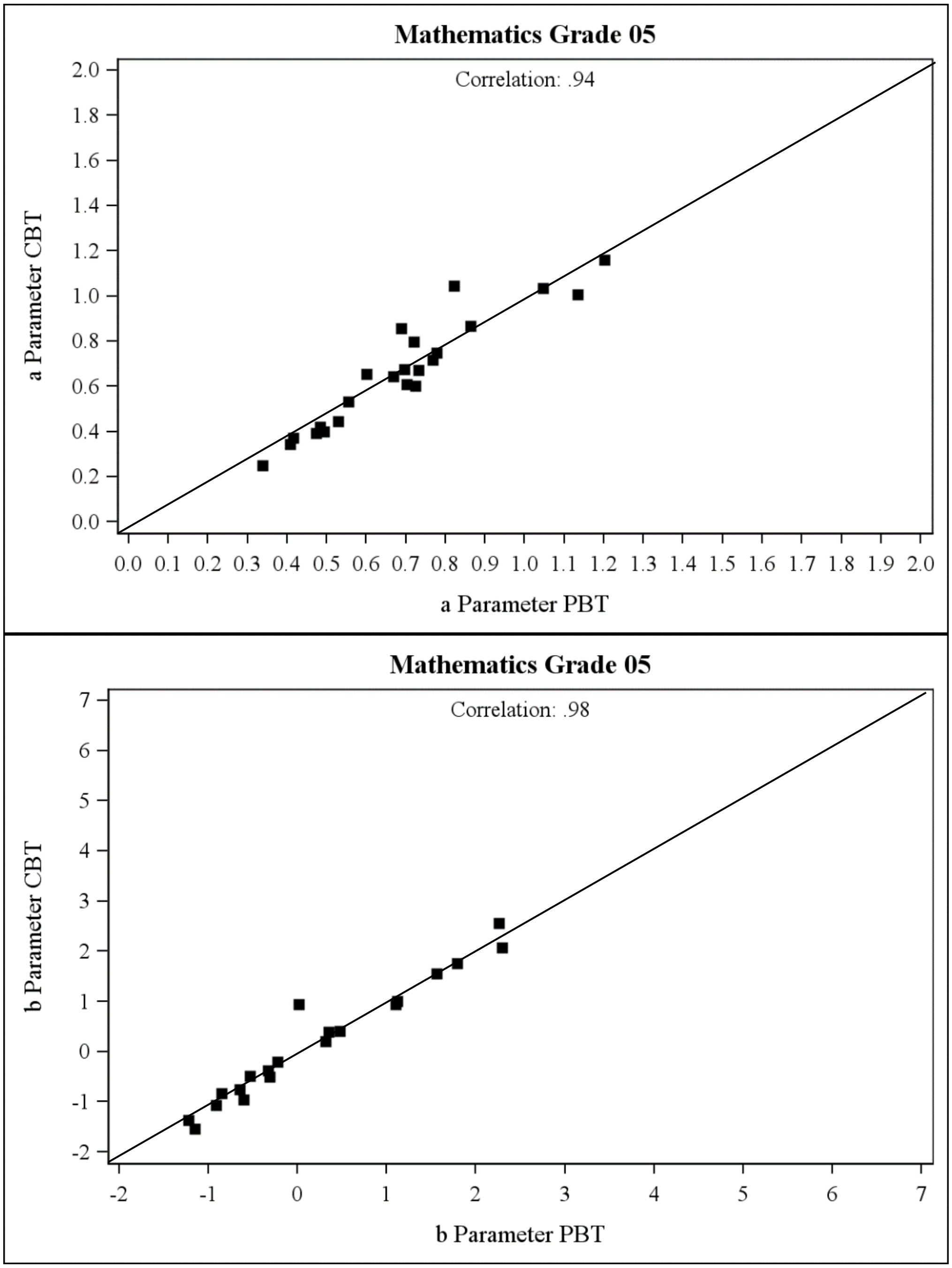


Figure 5.4. Correlation between Difficulty and Discrimination Parameter Estimates across Modes for Mathematics Grade 5.

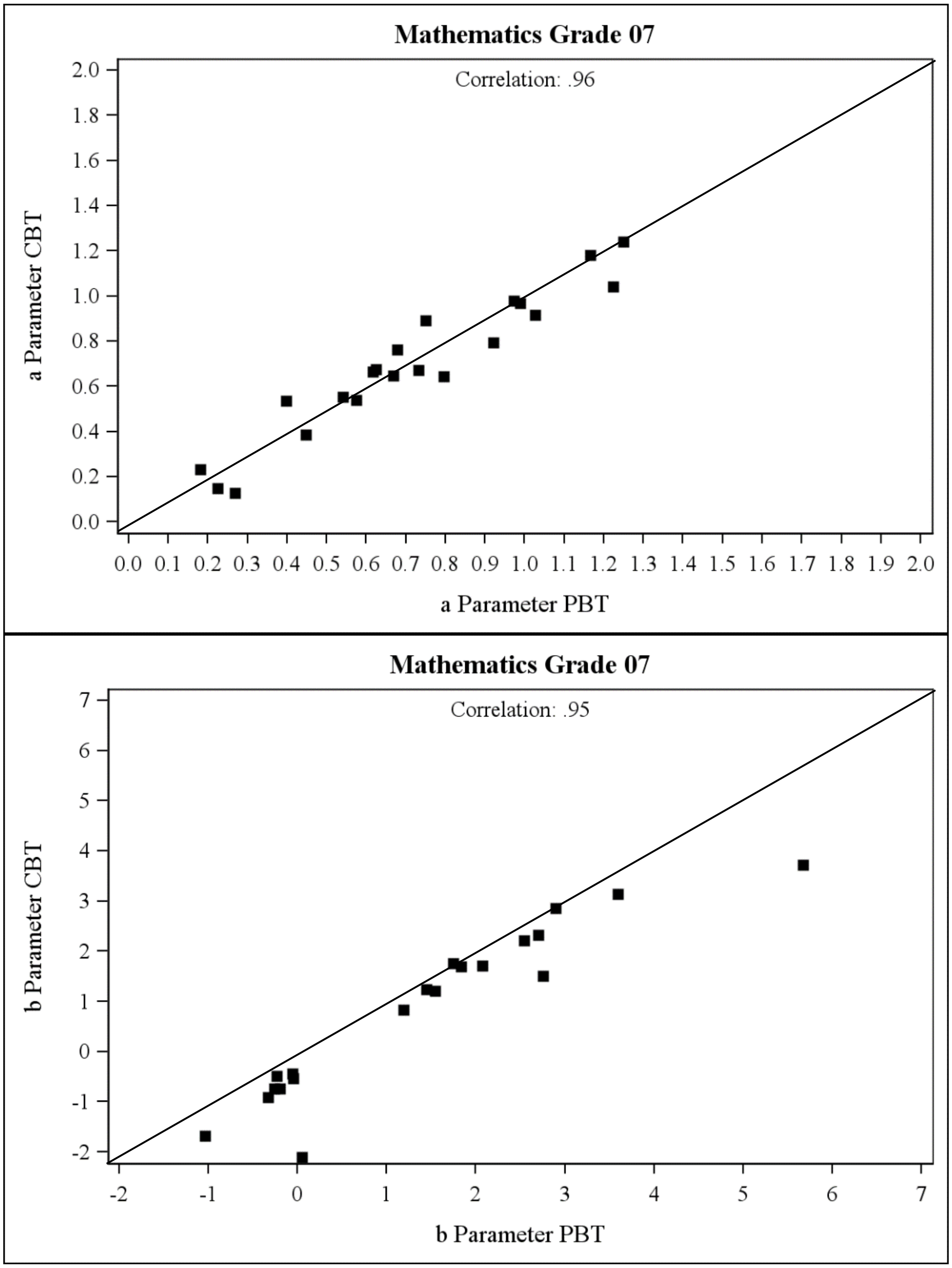
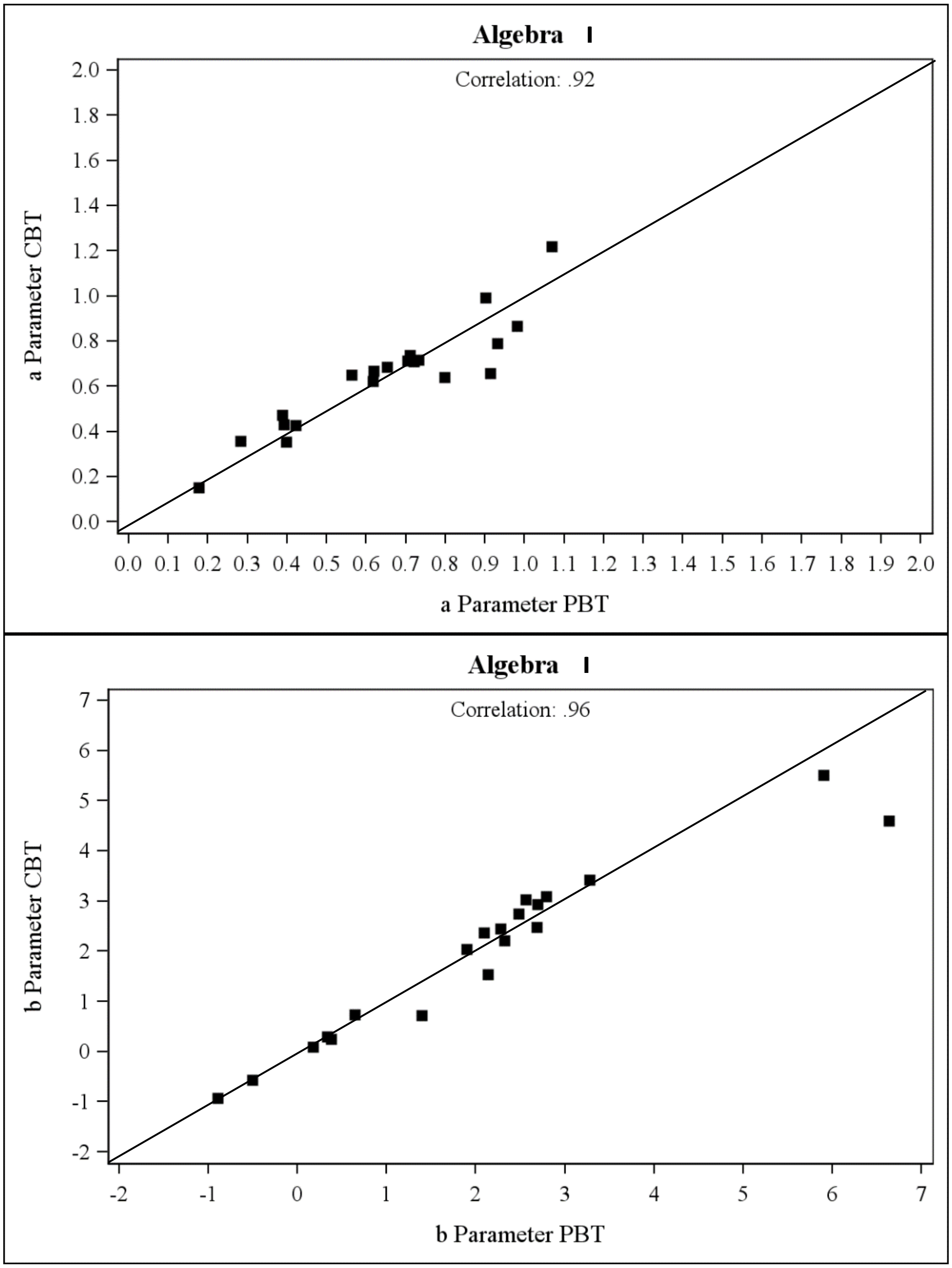


Figure 5.5. Correlation between Difficulty and Discrimination Parameter Estimates across Modes for Mathematics Grade 7.



Algebra I

Algebra I

Figure 5.6. Correlation between Difficulty and Discrimination Parameter Estimates across Modes for Algebra I.

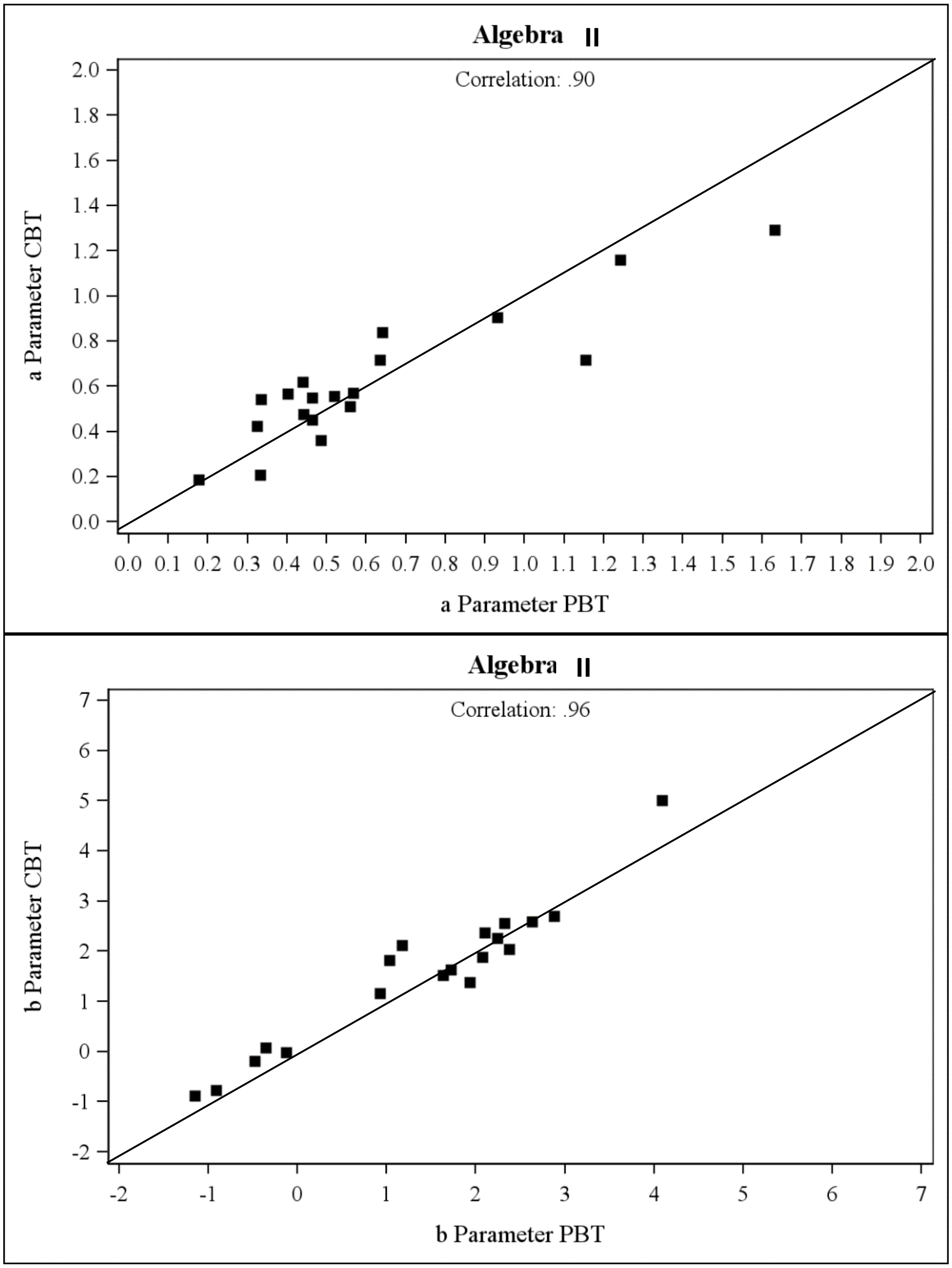


Figure 5.7. Correlation between Difficulty and Discrimination Parameter Estimates across Modes for Algebra II.

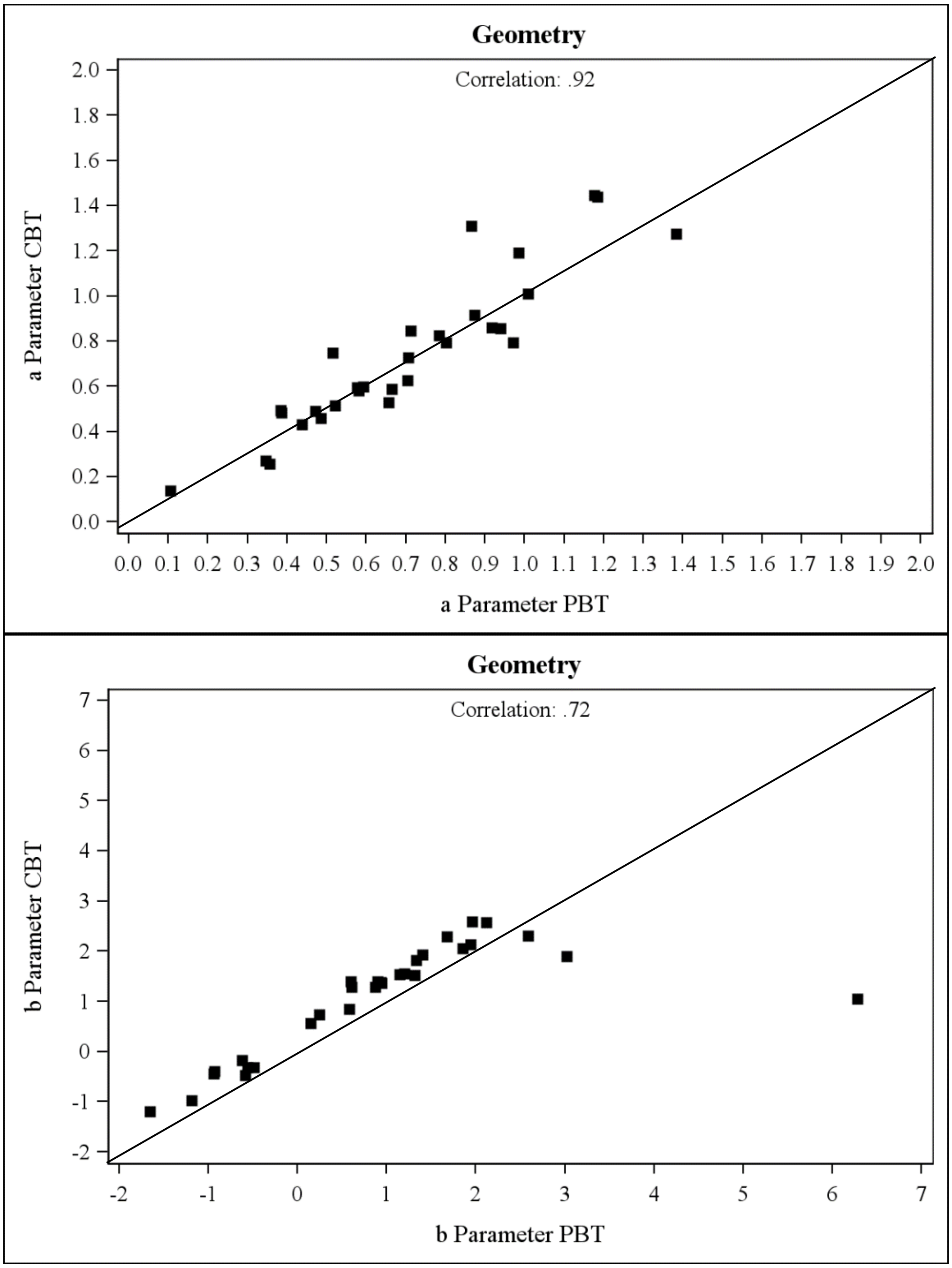


Figure 5.8. Correlation between Difficulty and Discrimination Parameter Estimates across Modes for Geometry.

Section 6: Analyses and Results Pertaining to the Similarity of Student Performance across Modes

6.1 Overview

The following sections summarize test statistics used to facilitate “test-level” comparisons across groups assessed using different test modes. These score comparisons involve the calculation of effect sizes, which characterize the importance of mode differences more directly than does statistical significance. All analyses were conducted on the sample obtained by propensity score matching for selected forms and grade levels.

6.2 Summary Test Statistics

6.2.1 Objectives

The goal of the analyses conducted in this section was to evaluate the degree to which test scores are comparable across modes. Particularly, are there substantial differences in reliability and average test scores for the overall test as well as for the common items? Substantial differences in test reliability could potentially impact the generalizability of PARCC assessment results to other situations or contexts. Differences in average test scores across modes might indicate potential construct-irrelevant variance that should be further investigated.

6.2.2 Method

The test score summaries include the raw score means and standard deviations and stratified alpha for common items shared between the form pairs. These summaries were also provided for all items (common and unique) associated with selected form pairs.

The stratified alpha reliabilities were computed for each selected form pair and grade level using the following formula:

$$_{strat\alpha}\rho = 1 - \frac{\sum s_{X_j}^2 (1 - \alpha_j)}{s_X^2} \qquad (4)$$

where, $s_{X_j}^2$ is the variance for stratum *j* of the test, s_X^2 is the total variance of the test, α_j is Cronbach’s alpha for stratum *j* of the test, and the summation is over the strata.

To compare reliability estimates across modes for all items appearing on a form pair, the stratified alpha estimates were adjusted using the Spearman-Brown formula. Specifically, since some test forms might have included items that could not be scored for various reasons, the overall test length could differ across modes. Therefore, the reliability estimates based on all item raw scores might not be comparable without an adjustment. The Spearman-Brown formula was used to adjust the all-item raw score reliabilities to the intended length of the assessment.

To summarize the relative performance at the test form level, means and standard deviations were computed. However, because many forms differed in numbers of items administered and scored, as well as in total points available, the scores were converted to percentages of total possible points within each form and means and standard deviations are reported in this metric. The relative performance of the common items between forms was summarized by providing the means, standard deviations, and effect sizes. The effect sizes were calculated for each pair of form level common item scores, in which groups differed by testing mode.

Effect size, *d* was computed as follows (Cohen, 1988, p. 20):

$$d = (M_{CBT} - M_{PBT}) / s \qquad (5)$$

where,

d is the effect size,

M_{CBT} is the mean of the common item scores for the CBT group,

M_{PBT} is the mean of the common item scores for the PBT group,

s is the pooled standard deviation of the CBT and PBT groups.

Besides raw score statistics, the scale scores were also summarized for both modes. The scale scores used in this report were obtained based on the scoring tables created after scaling between CBT and PBT forms in 2015 PARCC operational test equating. Detailed information about 2015 PARCC operational test equating and scale scores can be found in the *2015 PARCC Operational Test Technical Report* (PARCC, 2016).

In addition, a special investigation linking students' 2015 PARCC scores to their prior achievement was conducted. The first step in the study was to create matched groups of students taking tests in both modes. The propensity score matching approach was used to match students on a set of variables presumed to be relevant to PARCC test scores. However, due to the unavailability of most students' previous state assessment scores, this variable was not included in the matching process. One state (referred to as State S for confidentiality) did provide prior assessment scores, and it was used to further investigate potential performance differences due to mode. Specifically for State S, prior test scores were summarized on the sample of students derived from the propensity score matching process to determine whether their prior achievement was comparable across mode. Then poststratification was used to assign PBT students sampling weights based on their prior state test scores. Poststratification is widely used in survey analysis to account for underrepresented groups in the population. It is also useful for improving the precision of point estimators by minimizing bias. The basic technique is to divide the sample into strata and calculate the poststratification weights for each sample case in the strata (Holt & Smith, 1979; Little, 1993). Using every possible prior test score as the strata and the CBT students as the target population, poststratification weights for PBT students were computed. After weighting, the CBT and PBT students' prior test score distributions were similar. The weighted PBT students' PARCC test scores were calculated and compared to CBT students' scores. The goal of using students' prior test score poststratification weights to weight PARCC test score was to determine average PARCC test scores for students testing on paper if they had the same prior test score distribution as students testing online. The poststratification weights were calculated using the following formula:

$$w_s = \frac{n_{PBT} \times p_{CBT.s}}{n_{PBT.s}} \qquad (6)$$

where w_s is the poststratification weight for prior test score stratum *s*, $n_{PBT.s}$ is the number of PBT students in stratum *s*, $p_{CBT.s}$ is the proportion of CBT students in stratum *s*, and n_{PBT} is the total PBT sample size.

6.2.3 Results

Table 6.1 provides the test score summary for ELA/L for selected grades and form pairs by mode. The highest possible score for grade 3 ELA/L test was 100, therefore, the mean and standard deviation in percentage were the same as original statistics. For the all-item raw scores, PBT forms had higher mean raw scores in the percentage metric than CBT forms for grades 3 and 9. The grade 7 ELA/L CBT form had a higher mean in the percentage metric than the PBT form. The effect sizes associated with the common-item raw scores for grades 3 and 9 were large and negative (-.22 and -.30, respectively) but small and positive (.03) for grade 7. The reliabilities were comparable across modes for all-item and common-item raw scores. Because most of the items were common between modes, the all-item coefficient alphas were very similar to the common-item coefficient alphas, and there were minimal differences between modes.

Table 6.2 provides the test score summary for mathematics for selected grades and form pairs. For all subjects and grade levels, the PBT form had a higher mean total raw score in the percentage metric than the CBT form. The differences between modes in total raw scores in the percentage metric were larger for Geometry and Algebra II than for other subjects/grade levels. The effect sizes associated with the common-item raw scores are large for grade 7, Geometry and Algebra II, .33, -.41, -.34, respectively. As may be seen by the signs of these effect sizes, PBT students scored higher on common items than CBT students for Geometry and Algebra II, but CBT students scored higher for grade 7 Mathematics. The effect sizes associated with the common-item raw scores were small for grade 5 and Algebra I, -.02 and .06, respectively. The reliabilities were comparable across modes with trivial differences for all-item and common-item raw scores except for Geometry for all items and Algebra II for common items. For Geometry the coefficient alpha for the all-item scores was higher for the PBT form (.91) than for the CBT form (.87). For Algebra II the common-item alpha was lower for the CBT mode (.74) than for the PBT mode (.80).

Table 6.1 Test Raw Score Summary for ELA/L for Selected Grades and Forms by Test Mode

		Raw Score : All Items								Raw Score : Common Items						
Grade	Mode	#items	Min	Max	Mean	Mean in Percentage Metric	SD	SD in Percentage metric	Alpha	#items	Min	Max	Mean	SD	Alpha	Effect Size
3	CBT	37	2	91	36.11	36.11	18.44	18.44	.93	31	2	82	31.37	15.93	.92	-.22
	PBT	37	2	95	39.74	39.74	18.85	18.85	.93	31	1	84	34.98	17.01	.92	
7	CBT	50	4	127	53.75	39.23	24.79	18.10	.94	47	4	122	50.73	23.90	.94	.03
	PBT	50	3	127	51.72	37.75	26.07	19.03	.94	47	3	122	49.87	25.36	.94	
9	CBT	50	4	121	57.78	42.17	23.92	17.46	.94	44	3	111	52.87	21.75	.93	-.30
	PBT	50	3	126	66.31	48.40	23.65	17.26	.93	44	2	116	59.49	21.80	.92	

Table 6.2 Test Raw Score Summary for Mathematics for Selected Grades/Subjects and Forms by Test Mode

		Raw Score : All Items								Raw Score : Common Items						
Grade	Mode	#items	Min	Max	Mean	Mean in Percentage Metric	SD	SD in Percentage Metric	Alpha	#items	Min	Max	Mean	SD	Alpha	Effect Size
Mathematics 5	CBT	52	1	78	31.74	38.71	14.41	17.58	.93	23	0	30	14.18	5.58	.83	-.02
	PBT	52	4	81	34.81	42.46	16.01	19.53	.94	23	0	30	14.32	5.93	.84	
Mathematics 7	CBT	50	1	74	22.10	26.96	12.21	14.89	.92	21	0	33	9.84	5.32	.83	.33
	PBT	50	1	78	22.20	27.07	13.67	16.67	.92	21	0	33	8.09	5.29	.84	
Algebra I	CBT	53	3	80	20.01	20.62	11.54	11.89	.91	20	0	33	7.85	5.18	.81	.06
	PBT	53	2	85	21.60	22.26	11.88	12.25	.90	20	0	35	7.55	5.22	.81	
Geometry	CBT	53	2	85	24.90	25.67	15.70	16.19	.87	30	1	53	15.11	9.94	.89	-.41
	PBT	53	3	88	33.33	34.36	16.63	17.15	.91	30	1	54	19.47	10.61	.88	
Algebra II	CBT	52	2	53	19.76	19.56	10.76	10.65	.93	20	0	29	9.54	4.92	.74	-.34
	PBT	54	2	91	28.60	26.73	15.25	14.25	.93	20	0	38	11.45	6.17	.80	

Tables 6.3 summarizes the PARCC ELA/L and mathematics scale scores of propensity-score-matched students in for the entire analysis sample for all selected form pairs and grade levels. Effect sizes of mean scale score differences (CBT-PBT) between testing modes are also listed. The differences between CBT and PBT students' PARCC scale scores varied across different grades and subjects. For ELA/L grade 7, mathematics Grade 5, and Algebra I, the effect size are small with absolute values less than .10. Effect sizes are larger for the remaining subjects and grades with absolute values greater than .20. Among the subjects and grades with large effect sizes, only in grade 7 mathematics did CBT students have a higher scale score mean than PBT students. The effect sizes of scale scores are, in general, consistent with effect sizes of common item scores in Tables 6.1 and 6.2 except for Algebra II (-.34 for common items and -.20 for propensity-score-matched). As shown in Table B.2 Algebra II has the highest number of C-level DIF items (12 in total) which were later removed from operational linking between CBT and PBT, which may have caused the difference in effect sizes of common items and scale scores. A summary of PARCC scale scores for students from State S in the matched analysis sample are provided in Table 6.4. The presented results do not include Algebra II because no students in State S were administered the selected core form pairs in that subject. In addition, no results are presented for ELA/L grade 3 students because no prior state test scores were available. Thus no results were included in any of the subsequent analyses conducted using State S sample for these two subjects and grades. With the exception grade 9 ELA/L, the mode differences in the PARCC scale score means for the overall matched sample group are dissimilar to those for State S.

Table 6.5 provides summary statistics of State S students' prior achievement, which were test scores in a related content area. For example, Algebra I students' prior test scores are students' mathematics grade 7 and 8 state assessment scores. To preserve anonymity for State S and eliminate the scale differences of prior assessment scores from different grade levels, all prior state test scores were converted to *z*-scores prior to analysis. If the samples of State S students taking CBT and PBT tests are equivalent in their prior achievement, the effect sizes in Table 6.5 should match those in Table 6.4. The differences in effect sizes listed in Tables 6.4 and 6.5 indicates that the samples of State S CBT and PBT students after propensity score matching do not have comparable prior achievement which is one of the most important predictors of current PARCC scores. The differences in CBT and PBT students' performance shown in Table 6.4 were a result of mode effects or nonequivalent samples. Therefore, poststratification weights were used to calculate PBT students' summary statistics to adjust for prior achievement differences between PBT and CBT students. Table 6.6 provides a PARCC scale score summary of ELA/L and mathematics for all selected form pairs and grade levels after PBT students' scores were weighted by prior state test score post-stratification weights. After the adjustment, the differences between PBT and CBT students' scale scores shrunk significantly for mathematics grades 5 and 7 and Algebra I. Medium effect sizes were found for ELA/L grade 9 and Geometry favoring PBT students and a small effect sizes for ELA/L grade 7 also favoring PBT students.

Table 6.3 Summary of PARCC Scale Scores of ELA/L and Mathematics of Propensity Score Matched Students for Selected Form Pairs and Grade Levels

Subject/Grade	CBT			PBT			Effect Size
	N	Mean	SD	N	Mean	SD	
ELA/L							
3	5806	734.6	37.4	7978	743.3	40.1	-.22
7	5233	739.3	34.3	6558	737.8	36.7	.04
9	1548	745.7	33.7	2813	756.4	33.2	-.31
Mathematics							
5	5618	739.3	28.3	7513	737.5	30.6	.06
7	4008	736.4	25.8	7754	726.1	28.5	.37
Algebra I	2638	736.8	31.7	5475	734.4	33.5	.07
Geometry	868	737.6	26.7	857	749.8	26.4	-.45
Algebra II	1235	720.8	36.5	893	726.8	36.8	-.20

Table 6.4 Summary of PARCC Scale Scores of ELA/L and Mathematics of Propensity Score Matched Students from State S for Selected Form Pairs and Grade Levels

Subject/Grade	CBT			PBT			Effect Size
	N	Mean	SD	N	Mean	SD	
ELA/L							
7	1870	742.47	32.96	2715	737.79	34.94	.14
9	1049	747.84	31.63	1825	756.98	32.14	-.29
Mathematics							
5	1098	745.10	26.92	2428	741.95	30.01	.11
7	836	740.19	24.97	3160	725.48	26.89	.56
Algebra I	798	742.70	28.58	2732	735.96	33.68	.21
Geometry	244	754.49	20.22	410	761.91	19.08	-.38

Table 6.5 Summary of Prior State Test Scores of Related Subjects of Propensity Score Matched Students for Selected Form Pairs and Grade Levels

	CBT			PBT			Effect
Subject/Grade	N	Mean	SD	N	Mean	SD	Size
ELA/L							
7	1869	0.13	0.97	2711	-0.24	1.04	.36
9	1047	0.14	0.93	1822	0.07	0.98	.07
Mathematics							
5	1098	0.21	0.86	2428	0.13	0.98	.09
7	835	0.13	0.90	3156	-0.49	0.99	.61
Algebra I	798	0.23	0.93	2732	-0.11	1.06	.33
Geometry	244	0.98	0.94	410	1.01	0.74	-.04

Table 6.6 Summary of Weighted PARCC Scale Scores of ELA/L and Mathematics of Propensity Score Matched Students for Selected Form Pairs and Grade Levels

	CBT			PBT			Effect
Subject/Grade	N	Mean	SD	N	Mean	SD	Size
ELA/L							
7	1869	742.47	32.96	2711	747.78	33.82	-.16
9	1047	747.84	31.63	1822	758.99	31.38	-.35
Mathematics							
5	1098	745.10	26.92	2428	743.93	28.37	.04
7	835	740.19	24.97	3156	739.77	26.15	.02
Algebra I	798	742.70	28.58	2732	744.76	30.88	-.07
Geometry	244	754.49	20.22	410	761.57	20.17	-.35

6.2.4 Summary

Overall for both ELA/L and mathematics, there were differences in the percentages of points earned in favor of PBT for all subjects and grade levels assessed except for ELA/L grade 7. The common items raw score effect sizes are consistent with PARCC operational scale score effect sizes for all subjects and grade levels but Algebra II. An analysis of the State S student sample revealed that the CBT and PBT student samples were not comparable in prior achievement which can affect the comparison of PARCC scores between modes. After adjusting PBT students' prior test score distributions to match CBT students' prior distributions the differences in students' performance between modes decreased for mathematics in grades 5 and 7 and in Algebra I but not for other subjects and grades. Meanwhile, only two ELA/L grades were included in the study since no prior achievement scores were available for grade 3 students, therefore, results were mixed for ELA/L tests.

Section 7: Evaluation of Estimated Equating Functions

7.1 Overview

Section 7 compared the test characteristic curves (TCCs) of common items that appeared on both CBT and PBT tests. The differences between common item TCCs provides evidence to evaluate whether common items perform similarly across modes.

7.2 TCC Comparison of Common Items

7.2.1 Objectives

In the case of mode comparability, regardless of the whether the research indicates an underlying mode effect between items appearing in both modes (i.e., CBT and PBT), some degree of equating will be required since there are subsets of items that cannot be delivered in both modes (e.g., TEI item types can only be delivered in CBT). However, the quality of the equating results will largely be a function of the common items shared between the modes. If the common items perform similarly across modes, then the TCC for items appearing on PBT should look nearly identical to the test characteristic curve for items appearing on computer, after transformation. Moreover, the difference in the expected number correct scores given ability theta between the CBT and PBT modes should be within the difference that matters (DTM) criterion (Dorans & Feigenbaum, 1994). The DTM criterion is defined as any difference that would have an impact on the reported scores a test taker would receive after rounding. Differences exceeding a DTM of a half a point on the equated raw score scale are considered significant in this study.

7.2.2 Method

Consistent with the approach used for the operational analysis, two separate calibrations were conducted. The Stocking and Lord (1983) approach was used to place the item parameter estimates from the PBT form onto the scale of the computer form. Common items with C-level DIF (refer to Section 4.3.3 on DIF analysis), polyserial correlation less than 0.1 or weighted root mean square difference (WRMSD) exceeding specified criteria (refer to Table 7.1) were removed from the common item set before scaling. All common items were found to have acceptable WRMSD during scaling. Please refer to Table 7.2 for the number of C-DIF items and the number low polyserial items removed from each test form. After transforming the item parameter estimates onto a common scale, TCCs were established for each mode. The differences between the TCCs from the two modes were evaluated as a function of theta. Differences exceeding the difference that matters (DTM) criterion would serve as evidence for lack of comparability between the modes for the common items.

Table 7.1. Weighted Root Mean Square Difference Criteria for Removing Common Items

Categories	Points	WRMSD/points	WRMSD
2	1	.100	.100
3	2	.075	.150
4	3	.075	.225
5	4	.075	.300
6	5	.075	.375
7	6	.075	.450
>=8	>= 7	.090	.999

Table 7.2 Number of Items Removed from Common Item Set for ELA/L and Mathematics

Subject/Grade	Number of Items Removed for C-DIF	Number of Items Removed for Low Polyserial Correlation	Total Number of Items Removed
ELA/L 3	0	0	0
ELA/L 7	0	0	0
ELA/L 9	6	0	6
Mathematics Grade 5	2	0	2
Mathematics Grade 7	0	1	1
Algebra I	0	2	2
Geometry	1	3	3*
Algebra II	2	2	4

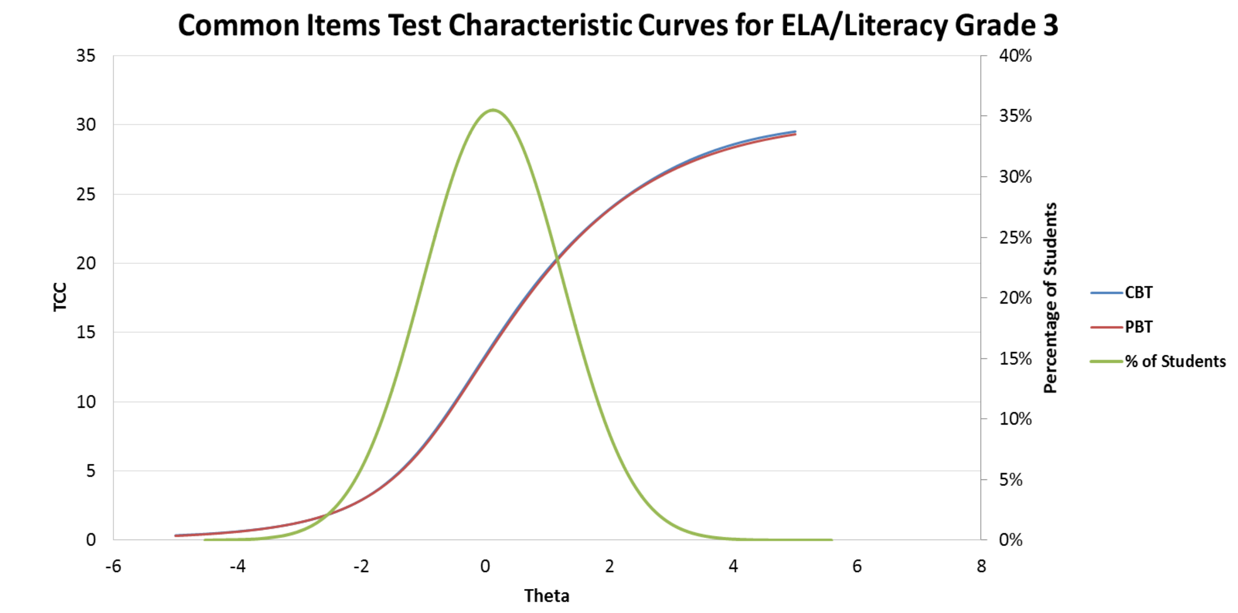
Note: *One item in Geometry was removed for having both C-level DIF and low polyserial correlation.

7.2.3 Results

Figures 7.1 through 7.8 provide comparisons between common item TCCs based on item parameter estimates from calibration and scaling of selected CBT and PBT forms for each chosen course/grade level for ELA/L and mathematics. The theta distribution of students in the analysis sample was also included in each graph on a separate axis on the right side of the graph; the TCC axis is shown on the left side. Due to the different numbers of common items and score points associated with common items, the TCC scales are not consistent across subjects and grade levels. To assist the visual evaluation, the minimum and maximum values of the TCC differences are also listed at the bottoms of the graphs. Differences exceeding the difference that matters (DTM) criterion (>0.5 raw score point) were found for several tests and the corresponding regions of the TCCs are marked by purple dots in the graphs to help identify these regions.

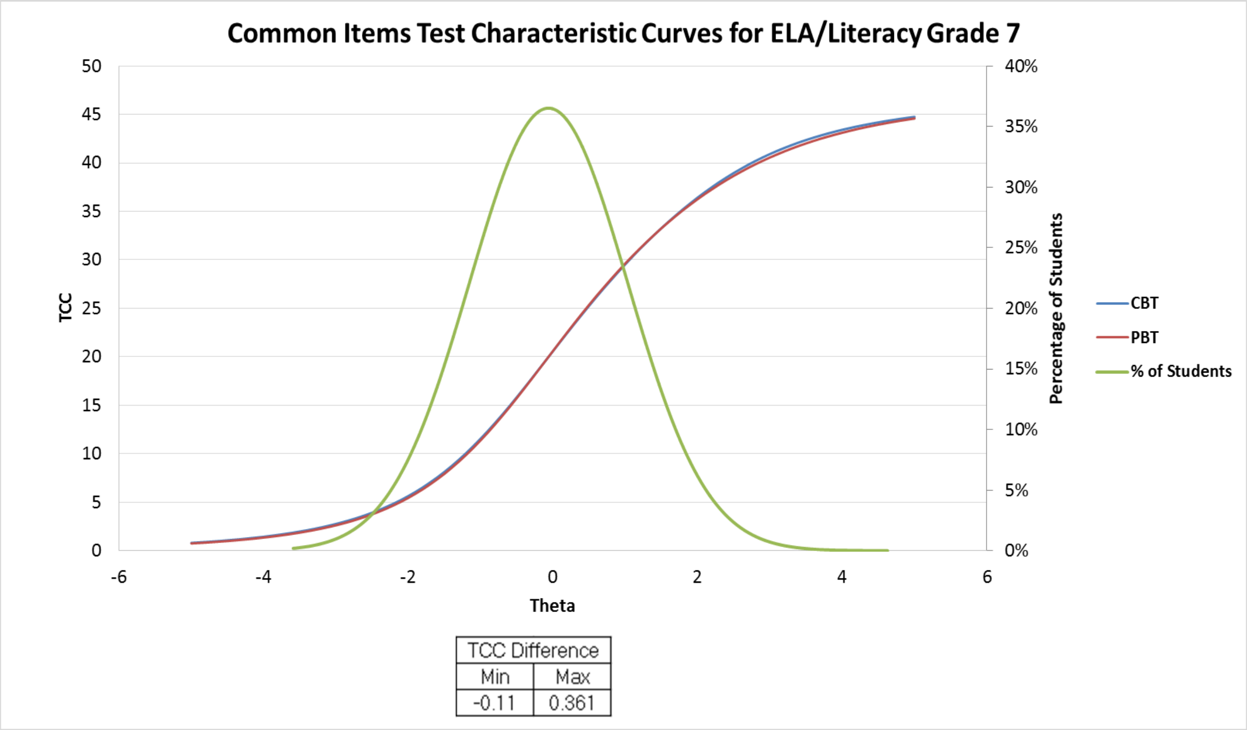
For ELA/L grades 3 and 7 the CBT and PBT TCCs are nearly identical with no TCC differences exceeding the DTM criterion. For ELA/L grade 9 TCC differences exceeding the DTM criterion occurred in two regions. Differences exceeding DTM occurred for theta values between 1.9 and 4.6, which represented 5% of students. However, there were a substantial number (18%) of students in the range for differences exceeding DTM for theta values between -2.1 and -0.8.

All TCC differences were within the DTM criterion for mathematics grade 5 and 7 and Algebra I. For Algebra II, the TCCs were within the DTM criterion for most of the theta scale with the exception of values between 2.5 and 4 where less than 1% of students were located. TCC differences exceeding the DTM criterion were found for Geometry for theta values between -2 and 0 where approximately 27% of students were located.



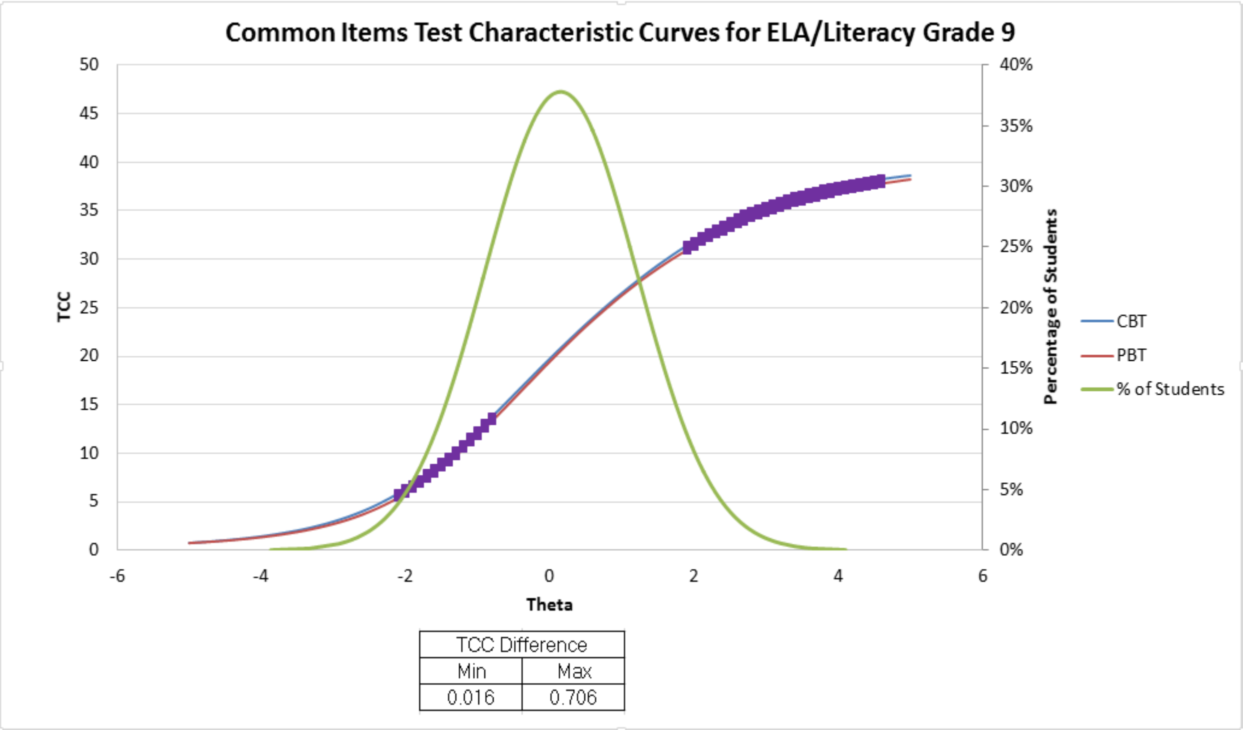
TCC Difference	
Min	Max
-0.007	0.22

Figure 7.1 Common Item TCCs for ELA/L Grade 3.



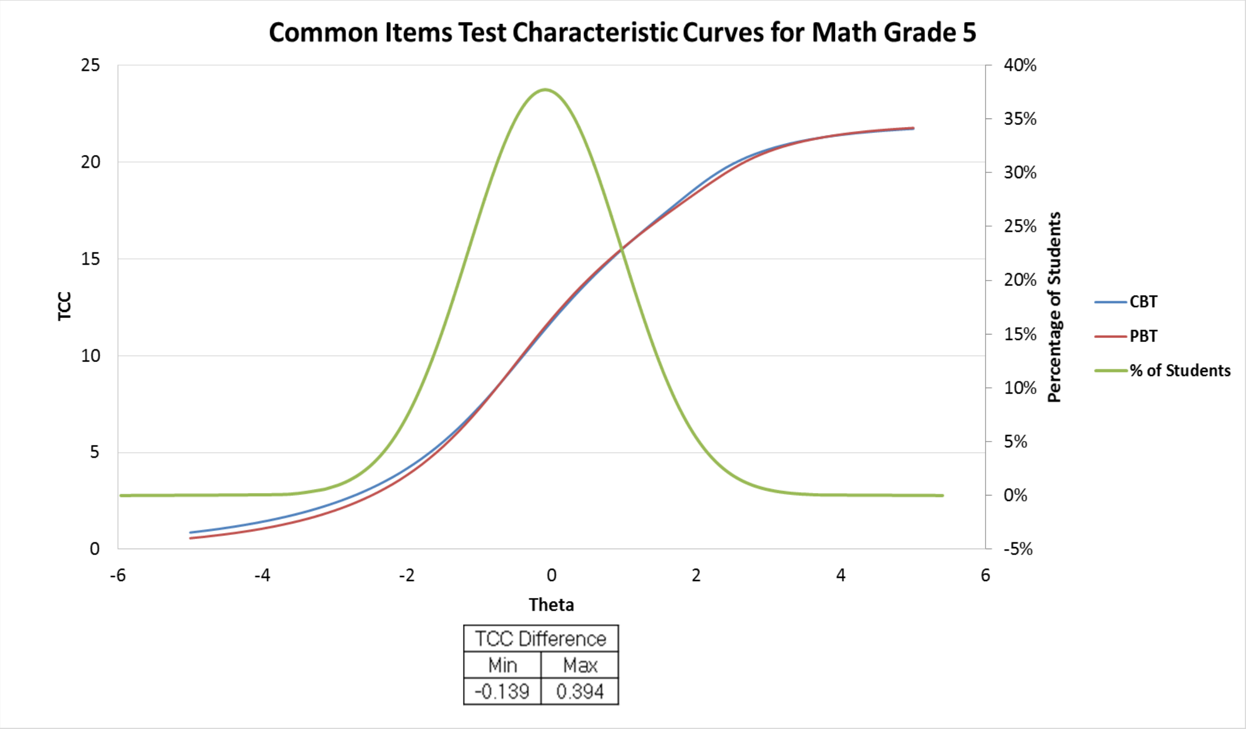
TCC Difference	
Min	Max
-0.11	0.361

Figure 7.2 Common Item TCCs for ELA/L Grade 7.



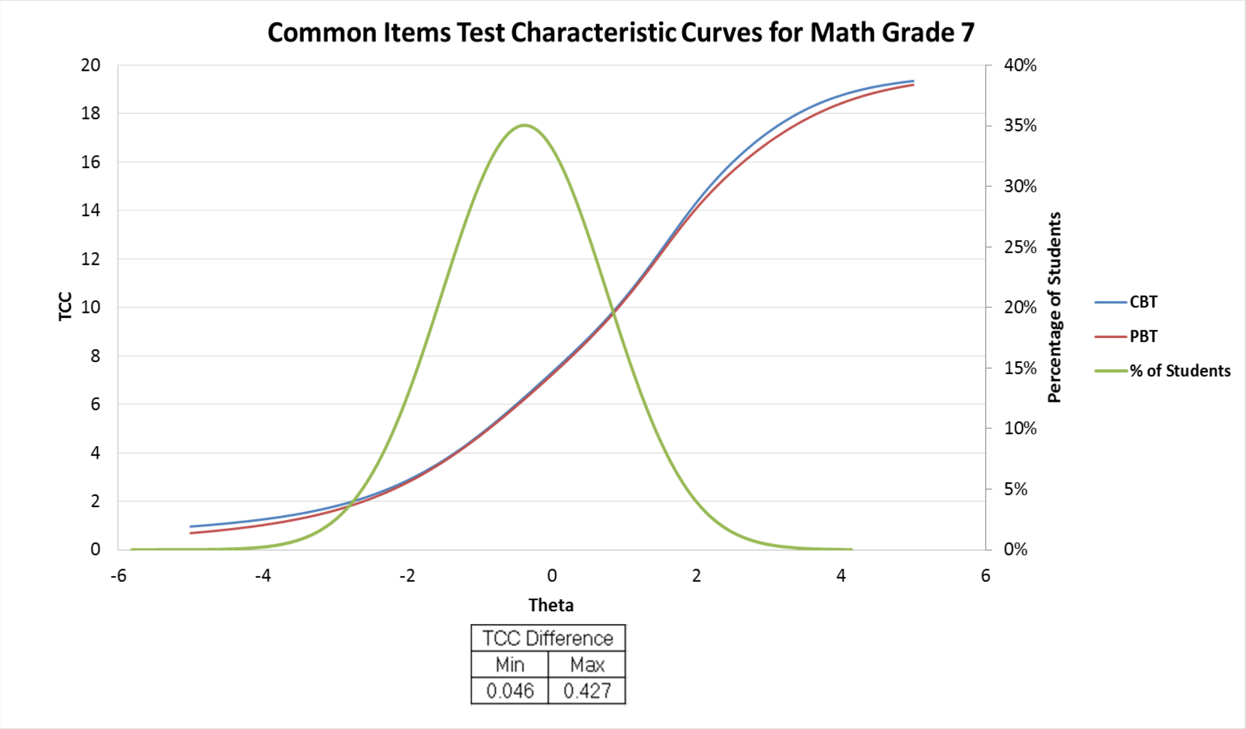
TCC Difference	
Min	Max
0.016	0.706

Figure 7.3 Common Item TCCs for ELA/L Grade 9.



TCC Difference	
Min	Max
-0.139	0.394

Figure 7.4 Common Item TCCs for Mathematics Grade 5.



TCC Difference	
Min	Max
0.046	0.427

Figure 7.5 Common Item TCCs for Mathematics Grade 7.

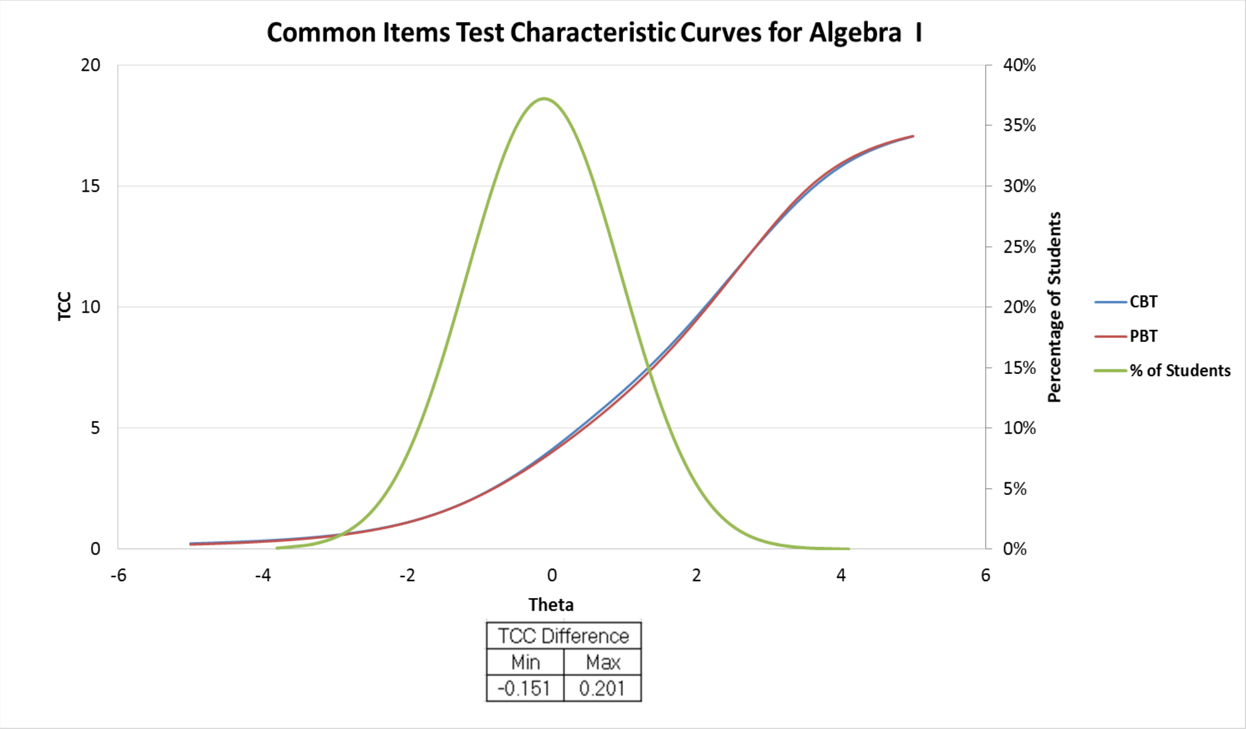


Figure 7.6 Common Item TCCs for Algebra I.

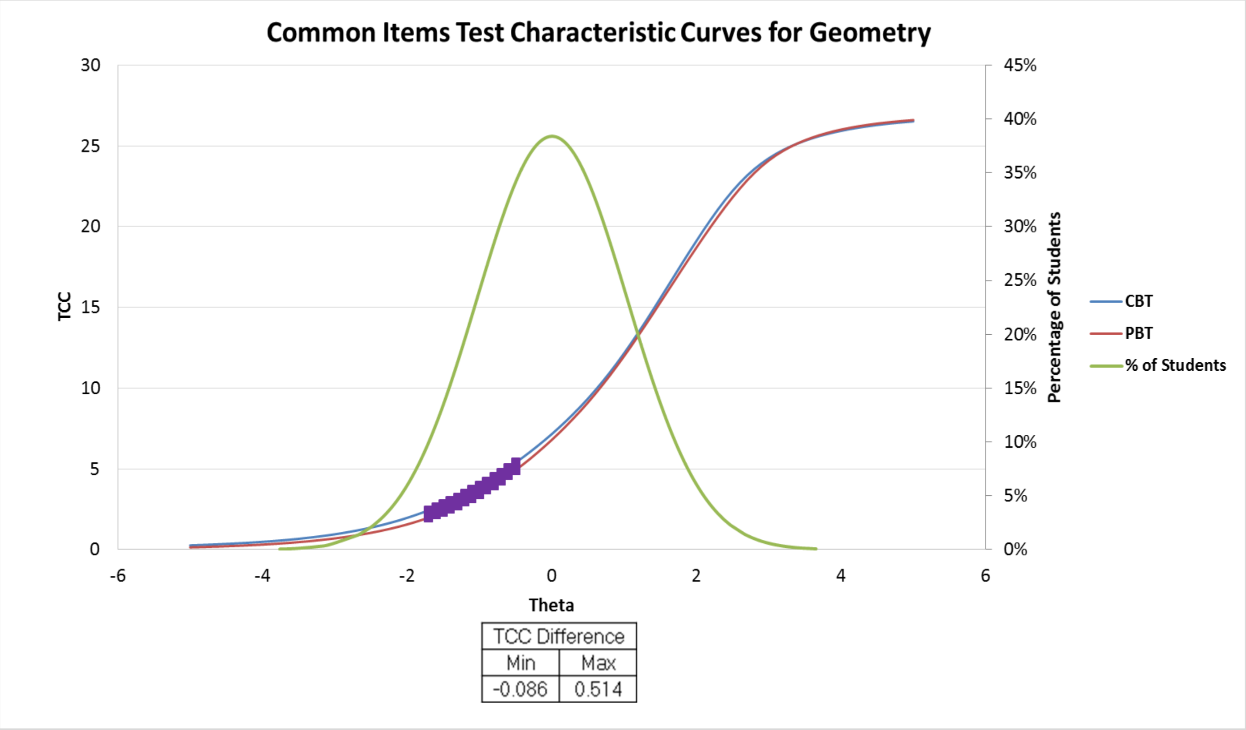
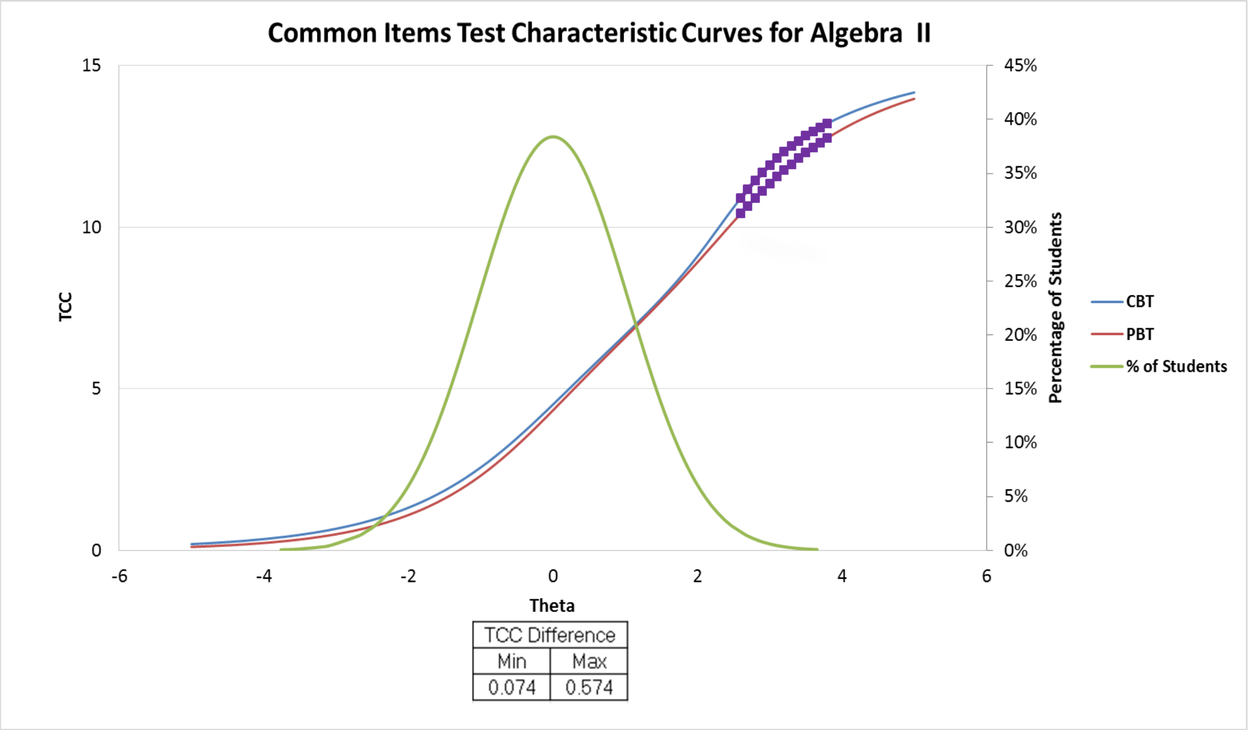


Figure 7.7 Common Item TCCs for Geometry.



TCC Difference	
Min	Max
0.074	0.574

Figure 7.8 Common Item TCCs for Algebra II.

7.2.4 Summary

Overall the differences between TCCs of different modes are small and within 0.5 raw score points for selected subjects/grade levels. Differences exceeding the DTM criterion were found in regions of the theta scale where large percentages of students are located for both ELA/L grade 9 and Geometry. These two tests also have the largest mode differences, in terms of effect sizes associated with PARCC scale scores after adjusting for prior achievement (refer to Table 6.6 in Section 6.2.3 for more information). Algebra II is another test with TCC differences larger than the DTM criterion. However, the differences were found in a region with very few students and thus had a minor impact on student scores as a function of test mode.

Section 8: Conclusions and Implications

8.1 Conclusions

The goal of the PARCC mode comparability study was to evaluate to what extent scores from the CBT and PBT form versions of the PARCC assessments can be considered comparable. The study focused on the following two questions:

1. Is the construct invariant between the two modes of test administration?
2. Is student performance similar between the two modes?

Several analyses were performed to address these research questions. Table 8.1 summarizes the analyses and findings including limitations of reported evidence from the analyses.

Table 8.1 Summary of Analyses, Findings, and Quality of Evidence from Mode Comparability Study

Research Questions	Analysis	Findings	Limitations of Evidence
Is the construct invariant between the two modes of test administration?	*z*-scores analysis	Item *z*-scores across modes were highly correlated except for Geometry Small differences in difficulties across mode except for PCR items for ELA/L and in Geometry. PCR items and Geometry items tended to be easier on PBT.	Group differences between CBT and PBT students could have impacted item statistics.
	DIF	Small percentage of items were flagged as C-DIF items for both ELA/L and mathematics;	Criterion scores may be affected by items functioning differently across modes.
	Item parameter correlation across modes in IRT	Item parameters from separate calibrations were highly correlated for nearly all subjects/grade levels and were largely robust to different calibration approaches.	Group differences between CBT and PBT students may have affected IRT estimation.
	Common item TCCs Differences between modes	Differences between common TCCs of different modes were always within 0.5 raw score points for selected subjects/grade levels except for ELA/L grade 9, Algebra II and Geometry	Group differences between CBT and PBT students may have affected IRT estimation.
Is student performance similar between the two modes?	Common items raw score effect sizes	The effect sizes were larger than 0.2 for half of the tests evaluated with directions inconsistent across subjects/grade levels	Raw score statistics were highly affected by CBT and PBT group differences.
	Scaled score effect sizes	Scale score effect sizes were consistent with common items effect sizes for all subjects and grade levels except Algebra II	Scaled score statistics were impacted by CBT and PBT group differences.
	Scaled score effect sizes after poststratification adjustment for State S	The PARCC scale score differences between modes were largely reduced for mathematics in grades 5 and 7 and in Algebra I but not for other subjects and grades. A significant mode effect in favor of the paper format was found for ELA/L grade 9 and Geometry assessments.	No contextual information about State S students' prior state assessment scores were available. Results based on one state and may not generalize to other PARCC states.

Before discussing the conclusions related to the two questions of interest, it is crucial to remember the limitations of the study discussed in Section 1.5. Due to the nature of the data collection design, students were not randomly assigned across modes (CBT and PBT). Schools/districts decided their students' testing mode and which students tested via CBT and PBT varied across states. For example, in one state approximately 98% of students tested online. In other states, approximately one-half of the students tested online. Attempts to use quasi-experimental adjustments based on the available demographic information were not adequate to generate matched samples equal in underlying ability. As a result, while differences were found across modes, it is not possible to definitively determine if these differences are real, an artifact of a mode effect, or some combination of both. While differences are larger in some grades and subjects than others, a consistent pattern of differences across modes was not observed from grade 3 through high school for either ELA/L or mathematics. Specifically, the direction of the mode effect often, but not always, favored PBT. Consequently, there is insufficient evidence to definitively answer the questions of interest in this mode comparability study.

The first question of interest was whether the construct was invariant between the two test modes. Correlations between modes for transformed item difficulties (*z*-scores) were in general very high except in Geometry. There were marginal differences (0.02 to 0.09) in the median difficulties across the assessments. Items appearing on the PBT forms tended to be easier (with higher *p-value*s) than on computer for all selected forms and selected subjects/grade levels. For ELA/L assessments, PCR trait items showed larger effect sizes of *p-value* differences than other items for all selected grade levels. In addition, for ELA/L assessments, extended text interaction items and items of lower cognitive complexity tended to have larger performance differences across modes. For mathematics, there were more constructed response items, fill-in-the-blank items, as well as items with high cognitive complexity performing differently across modes. The 2014 field test mode comparability study results found more fill-in-the-blank items flagged for *p-value* differences which was consistent with the current study. There were, however, more flagged items with low cognitive complexity which was different from the current study. A small percentage of items were flagged for performing differently across modes after conditioning on test taker ability for all assessments. Mathematics items had larger flagging rates (the maximum observed across grade levels was 15%) than ELA/L (the maximum observed across grade levels was 3%). About half of the ELA/L items flagged for C-Level DIF favored PBT while the other half favored CBT. For mathematics, in grades 3 through 8 more flagged items favored CBT than PBT whereas for high school more items favor PBT than CBT. Similar to findings from the 2014 field test mode comparability study, mathematics items that performed differently across modes tended to be fill-in-the-blank items, multiple choice multiple select items, text entry interaction items, and judgment items.

The analysis of IRT parameter estimates revealed that IRT difficulties and discriminations estimated separately within mode were highly correlated for nearly all grade levels and assessments. For grade levels with lower correlations between modes, removing items with outlier parameter estimates provided substantial improvement in correlation. Based on the recommended correlation thresholds for linking items from separate calibrations onto a common scale, linking PBT items onto the same scale as CBT items should not present problems at most grade levels. For grades and subjects, where the correlation thresholds are not met, additional refinements to the common item set, such as removing outlier items, could be performed to support linking. There were differences in IRT difficulties for the majority of grade levels, especially for ELA/L grade 3 and grade 9, mathematics grade 7, Algebra II and Geometry, which were consistent with item *p-value* differences in classical item analyses results. The IRT parameter estimates were largely robust to different calibration approaches. Overall the differences between common TCCs of different modes were small and within 0.5 raw score points for selected forms and subjects/grade levels. However, there were two subjects/grade levels (ELA/L grade 9 and Geometry) in which TCC differences exceeded the DTM criterion in regions of the theta scale where large percentages of students were located.

The high correlations for item difficulties (based on both classical test theory and IRT-based methods) and IRT-based discriminations between CBT and PBT items, in addition to very few item being flagged for DIF, provided strong evidence in support of construct invariance for the subjects/grades evaluated with the exception of Geometry and the PCR items for ELA/L. However, because there were no “sister forms” and there was significant variation with respect to the number of common items across CBT and PBT forms, more compelling construct validity evidence could not be obtained by conducting confirmatory factor analysis (CFA) to evaluate the factor structure across testing modes.

The second question addressed whether student performance was similar across the two modes. As shown in Tables 6.1 and 6.2, effect sizes of common item raw scores varied across grades in terms of both magnitude and direction. When comparing the performance on the common items, the effect sizes were larger than 0.2 for half of the tests evaluated. The scaled score effect size patterns were consistent with common item raw score effect sizes for most grade levels and subjects except for Algebra II (see Table 6.3).

Since propensity score matching, used to establish comparable groups across all PARCC states, was based solely on demographic information and only explained between 13 and 29 percent of the variation in PARCC scores, a supplemental analysis was conducted using data from the only state that provided prior state assessment data. The analysis of this state’s data revealed that the CBT and PBT student samples were not comparable in prior achievement for some of the subjects/grade levels. As shown in Table 6.6, after adjusting PBT students’ prior test score distributions to match CBT students’ prior distributions for State S, the PARCC scale score differences between modes were reduced for mathematics in grades 5 and 7 and in Algebra I but not for the other subjects and grades. Overall, there appears to be a significant mode effect in favor of the paper format for ELA/L grade 9 and Geometry assessments. The analysis of ELA/L claim scale scores indicated that effect sizes of writing scale scores were much larger than reading scale scores and favored PBT for grades 7 and 9.

It was not possible to ascertain whether student performance was similar between the two modes given unavailability of prior test scores from all participating PARCC states to support matching student samples based on ability. For the one state that provided prior test scores, there was evidence of score comparability for some but not all subjects. The results of this one state cannot be generalized to all PARCC states.

8.2 Implications

Interpretation of current study results was limited by the data collection design and unavailability of prior achievement scores. Differences found in item statistics and student’s performance may be a result of mode effects or noncomparability of PBT and CBT students. Analyses of State S students’ scores used poststratification to minimize group differences in terms of prior achievement but the results cannot be generalized to other PARCC states.

One significant implication of this study was that score comparability was inconsistent across content domains and grade levels. For State S there is evidence of score comparability in mathematics for grade 5, grade 7 and Algebra I, based on the examination of effect sizes. However, for other grades, particularly ELA/L grade 9 and Geometry, there were substantial differences in scores across mode.

Although a subset of grade levels were evaluated in this study, a thorough investigation of the items that were flagged for either *z*-score differences or for DIF is warranted. Specifically, content experts might be able to study cognitive processes required to answer these items for both modes, which might lead to some explanations of the differences. Such feedback could further aid item development as well as refine form assembly guidelines.

The current psychometric scaling procedure calls for separate calibrations of CBT and PBT items within each grade level. The PBT items were then linked to the CBT scale based on common items appearing in both modes. The recommendation from the PARCC field test mode comparability study was to exclude all between-mode C-DIF items from the linking sets. In essence, these C-DIF items would be treated as different items with their own unique item statistics by mode. The exclusion of C-DIF items might not be sufficient to ensure score comparability. For example, Geometry, which had large effect size differences in scale scores, also had a large number of B-DIF items favoring PBT. For mathematics tests, the criteria for removing items from common set should be expanded to include B-DIF items if results indicates a disproportionate share of these flagged items favoring one particular administration mode.

The issue is more complex for ELA/L and might require more than expanding the scaling criteria to exclude B- and C-DIF items. In evaluating the B-DIF items, nearly all such items were PCR trait items. Since all these items favored PBT, its cumulative effect might distort the scaling results. Additionally, in evaluating differences in average item scores between the modes for the pool of items used to link PBT to the CBT scale, the differences were nearly always minimized when all PCR trait items were excluded. Based on this result, two potential options should be considered: (1) all PCR trait items should be removed from the scaling process; (2) more focused attention should be given to human scoring of PCR trait items across mode.

Implementing the first option might create other issues in linking the CBT and PBT forms since the linking set would not be representative of the full test if all PCR trait items were removed. Regarding the second option, prior research has indicated that handwritten responses may earn higher scores than computer responses across equivalent samples (Bridgeman & Cooper, 1998; Powers, Fowles, Farnum, & Ramsey, 1994; Russell & Tao, 2004). Part of the observed score differences in constructed response items is attributed to differential expectations for handwritten and computer responses (Russell & Tao, 2004). An analysis of human scoring applied to paper and computer responses might shed light on whether this effect may have contributed to differences as compared to other factors (e.g., length of written versus typed responses). Regardless of whether explicit analyses are conducted along these lines, the results of this research suggest that a review of the training procedures with respect to sensitizing raters to the potential biases that may occur in evaluating handwritten versus computer responses is warranted.

8.3 Limitations

There were several notable factors that may have impacted the study findings. First, there was no random assignment across modes (CBT and PBT). Schools/districts decided their students' testing mode. The resulting student samples in different testing modes were likely not randomly equivalent. In order to conduct mode comparability analyses on the operational data, propensity score matching was used to create matched samples of CBT and PBT students. Ideally, samples should be matched on individual student test scores on previous state assessments. Due to the unavailability of state assessment data from most participating PARCC states at the time of analysis, however, the study only included demographic variables when matching CBT and PBT students. Furthermore, the severe differences in PBT and CBT student distributions on several matching variables created difficulty in selecting matching samples. In order to obtain samples of sufficient size to perform all analyses, a less stringent criteria

(Cohen’s *d* less than 0.2) was used to evaluate the matching results. The goal of well-matched samples cannot be achieved if the distributions of the two groups disparate. Future mode comparability studies should not rely on propensity score matching solely to obtain comparable CBT and PBT groups if online and paper students differ vastly on within state covariates.

For the one state (State S) that provided prior achievement data, its samples suggested that demographic variables only explained 13% to 29% of the variance in PARCC scale scores while prior state assessment scores that were related to a particular subject accounted for 53% to 66% of variance across selected subjects/grade levels. Analysis of State S students’ prior achievement data also found that the CBT and PBT samples created by PSM may not be comparable in their prior state test scores, which would have impacted the analyses results of current study. Prior achievement differences indicate students’ ability differed across modes. The effect sizes of *p-value* differences, *z*-score analyses, and raw and scale score comparisons based on PSM samples are likely confounded with students’ ability differences across modes. The sample differences might have less impact on DIF analyses since the DIF statistics were calculated conditionally on students’ ability.

Secondly, theta estimates from IRT scaling of CBT and PBT forms before removing C-DIF items were used as criterion scores for DIF, which may be affected by items functioning differently across modes, especially for ELA/L where PCR trait items have much more weight than other items.

The analysis of State S students’ PARCC scale scores provided some preliminary results of students’ performance across mode if CBT and PBT students were matched on prior achievement. However, since no contextual information about State S students’ prior state assessment scores were available, there may exist confounding effects that might have impacted the results. In addition, the post sampling analyses were conducted only on States S data and may not generalize to other states.

The current study was not conducted on all PARCC tests but on selected forms of certain grade levels and subjects. The results varied across grade levels and subjects, which suggests that any preliminary and descriptive conclusions based on these selected tests cannot be generalized to the tests that were not included in this study.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Austin, P. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, *27*, 2037-2049.

Beguin, A. A., & Hanson, B. A. (2001, April). Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating. Paper presented at the Annual Meeting of the National Council of Measurement in Education, Seattle. (Retrieved November 12, 2015 from http://www.cito.com/research_and_development/psychometrics/~/media/cito_com/research_and_development/publications/cito_report01_2.ashx).

Beguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000, April). Effect of multidimensionality on separate and concurrent estimation in IRT equating. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans. (Retrieved November 12, 2015 from http://www.b-a-h.com/papers/paper0002.html).

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment, 1*(1). (Retrieved August 4, 2003, from http://www.bc.edu/research/intasc/jtla/journal/v1n1.shtml).

Bennett, R. E. (2003). *Online assessment and the comparability of score meaning*. Educational Testing Service, Princeton, NJ. (Retrieved January 23, 2009 from http://www.ets.org/Media/Research/pdf/RM-03-05-Bennett.pdf).

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 November 12, 2015 from http://files.eric.ed.gov/fulltext/EJ838621.pdf).

Bridgeman, B. & Cooper, P. (1998). Comparability of scores on word-processed and handwritten essays on the Graduate Management Admissions Test. Paper presented at the annual meeting of the American Educational Research Association. San Diego, CA. (Retrieved September 28, 2009 from http://files.eric.ed.gov/fulltext/ED421528.pdf).

Brown, T., Chen, J., Ali, U., Costanzo, K., Chung, S., & Ling, G. (2015). *Mode Comparability Study Based on Spring 2014 Field Test Data*. Unpublished manuscript. Washington, DC: Partnership for Assessment of Readiness of College and Careers.

Choi, S. W., & Tinkler, T. (2002). Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coon, C., McLeod, L., & Thissen, D. (2002). *NCCATS update: Comparability results of paper and computer forms of the North Carolina End-of-Grade Tests* (RTI Project No. 08486.001). Raleigh, NC: North Carolina Department of Public Instruction.

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT.* (ETS Research Memorandum 94-10). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach*. (Research Report No. 91-47). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2002). *ETS standards for quality and fairness.* Princeton, NJ: Author.

Johnson, M. & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, and Assessment, 4*(*5).* (Retrieved November 3, 2015 from http://files.eric.ed.gov/fulltext/EJ843854.pdf).

Hanson, B.A. and Beguin, A.A. (2002). Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate Versus Concurrent Estimation in the Common-Item Equating Design. *Applied Psychological Measurement, 26* (1), 3-24.

Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics, 15,* 609–627.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15,* 234–249.

Holt, D., and Smith, T. M. F. (1979). Post stratification. *Journal of the Royal Statistical Society A, 142*, 33-46.

Keng, L., McClarty, K. L., & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas Assessment of Knowledge and Skills. *Applied Measurement in Education, 21(3),* 207-226. (Retrieved March 14, 2014 from http://dx.doi.org/10.1080/08957340802161774).

Kim, D. & 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 November 27, 2008 from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1634/1478).

Kim, S., & Kolen, M. J. (2004). STUIRT: A computer program for scale transformation under unidimensional item response theory models [computer software]. Retrieved from https://www.education.uiowa.edu/centers/casma/computer-programs.

Kim, S., Kolen, M. J. (2007). Effects on Scale Linking of Different Definitions of Criterion Functions for the IRT Characteristic Curve Methods. *Journal of Educational and Behavioral Statistics, 32*(4), 371-397.

Little, R. J. A. (1993). Post-stratification: A modeler's perspective*. Journal of the American Statistical Association, 88*, 1001-1012.

Lottridge, S. M., Nicewander, W. A., & Mitzel, H. C. (2010). Summary of the online comparability studies for North Carolina’s End-of-Course assessment program. In Winter, P. C., Ed. (2010). *Evaluating the Comparability of Scores from Achievement Test Variations*, Chapter 2. Washington DC: Council of Chief State School Officers. (Retrieved: April 2, 2013 from http://dpi.state.nc.us/docs/acre/assessment/onlinecompstudy.pdf).

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Muraki, E., & Bock, R. D. (2003). *PARSCALE: IRT item analysis and test scoring for rating-scale data* (Version 4.1). Chicago, IL: Scientific Software International.

Olson, L. (2003). Legal twists, digital turns: Computerized testing feels the impact of "No Child Left Behind." *Education Week, 12*(35), 11-14, 16.

Oregon Department of Education. (2007). Comparability of student scores obtained from paper and computer administrations. (Retrieved: March 14, 2014 from http://www.ode.state.or.us/teachlearn/testing/manuals/2007/doc4.1comparabilitytesatopandp.pdf).

Paek, P. (2005). Recent Trends in Comparability Studies. Pearson Educational Measurement, Iowa City, IA. (Retrieved April 2, 2013 from http://images.pearsonassessments.com/images/tmrs/tmrs_rg/TrendsCompStudies.pdf?WT.mc_id=TMRS_Recent_Trends_in_Comparability_Studies).

Partnership for Assessment of Readiness of College and Careers. (2016). *Technical Report for 2015 Administration*. Unpublished manuscript. Washington, DC: Author.

Partnership for Assessment of Readiness of College and Careers. (2015). *PARCC 2014 Field Test Technical Report*. Unpublished manuscript. Washington, DC: Author.

Parsons, L. (2001). Reducing Bias in a Propensity Score Matched-Pair Sample Using Greedy Matching Techniques. Proceedings of the Twenty-Sixth Annual SAS Users Group International Conference, Long Beach, CA. (Retrieved November 3, 2015 from http://www2.sas.com/proceedings/sugi26/p214-26.pdf).

Pearson Educational Measurement (2010). Appendix G: Comparability study of paper and pencil, and online administration of the Mod-MSA. (Retrieved: March 14, 2014, http://www.marylandpublicschools.org/NR/rdonlyres/E865B914-1C2D-4B39-A276-FBC02765E950/28802/2010_MOD_Math_TechReport_041411_APPENDIX_G.pdf).

Pearson Educational Measurement (2012). Mathematics Minnesota Comprehensive Assessment – Series III (MCA-III), Mode Comparability Study Report. (retrieved: April 2, 2013, from http://www.google.com/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=17&ved=0CFUQFjAGOAo&url=http%3A%2F%2Feducation.state.mn.us%2Fmdeprod%2Fidcplg%3FIdcService%3DGET_FILE%26dDocName%3D042214%26RevisionSelectionMethod%3DlatestReleased%26Rendition%3Dprimary&ei=ayIjU-jWNMb00gHG5oHYCQ&usg=AFQjCNH6sM19iUDaUbEJ_svK6XsVP4OYpA).

Poggio, J., Glasnapp, D., Yang, X., Beauchamp, A., & Dunham, M. (2005). Moving from paper and pencil to online testing: Findings from a state large scale assessment program. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, PQ, Canada.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment, 3*(6). (Retrieved August 19, 2006 from https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxwYXBlcnZlcnN1c3NjcmVlbnxneDozN2UyZDk2Y2RiZGFjZjRj).

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2*(6). (Retrieved November 3, 2015 from https://www.google.com/url?url=https://ejournals.bc.edu/ojs/index.php/jtla/article/download/1666/1508&rct=j&frm=1&q=&esrc=s&sa=U&ved=0CBQQFjAAahUKEwi3maLK0IvJAhXK5iYKHfhPAL0&usg=AFQjCNFhNlWmns9mNrN41F15_MyLvaGMPQ).

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31*(3), 220-233. (Retrieved September 28, 2009 from http://www.jstor.org/stable/1435267).

Randall, J., Sireci, S., Li, X., & Kaira, L. (2012). Evaluating the comparability of paper- and computer-based science tests across sex and SES subgroups. *Educational Measurement: Issues and Practice, 31*(4), *2- 12.*

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84,* 1024-1032.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2,* 169–188.

Russell, M. and Tao, W. (2004). Effects of handwriting and computer-print on composition scores: A follow-up to Powers, Fowles, Farnum, & Ramsey. *Practical Assessment Research & Evaluation, 9*(1*)*. Retrieved September 28, 2009 from http://PAREonline.net/getvn.sap?v=9&n=1.

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-Based Testing and Validity: A Look Back Into the Future. *Assessment in Education, 10*(3), 279-294.

Shadish, W. R., Clark, M. H., & Steiner, P.M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103,* 1334–1356.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75,* 417-453.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Wan, L., Keng, L., McClarty, K., & Davis, L. (2009). Methods of comparability studies for computerized and paper-and-pencil tests. *Teb st, Measurement & Research Services Bulletin*. Person Educational Measurement, Iowa City, IA. (Retrieved November 3, 2015 from http://images.pearsonassessments.com/images/tmrs/tmrs_rg/Bulletin_10.pdf?WT.mc_id=TMRS_Bulletin_10_Methods_of_Comparability_Studies).

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA (retrieved: May 22, 2013, http://images.pearsonassessments.com/images/tmrs/Score_Comparability_of_Online_and_Paper_Administrations_of_TAKS_03_26_06_final.pdf).

Way, W. D., Lin C., & Kong, J. (2008). Maintaining score equivalence as tests transition online: Issues, approaches and trends. Paper presented at the annual meeting of the National Council on Measurement in Education. (Retrieved November 3, 2015 from http://images.pearsonassessments.com/images/tmrs/Maintaining_Score_Equivalence_as_Tests_Transition_Online.pdf).

Appendix A

Table A.1 Characteristics of Items Flagged for Flagged for Effect Sizes for ELA/L Selected Forms and Grade Levels

	Flagged Effect Size Items		All Items	
Cognitive Complexity	Count	Percentage	Count	Percentage
Low	13	72	25	20
Medium	1	6	41	34
High	4	22	56	46
Response Type	Count	Percentage	Count	Percentage
Multiple: Multiple	0	0	3	2
Multiple: Single	0	0	3	2
Single: Multiple	1	6	17	14
Single: Single:	1	6	75	61
Single: Single: Single	8	44	8	7
Single: Single: Single: Single	8	44	16	13
Interaction Type	Count	Percentage	Count	Percentage
Choice Interaction	1	0	49	40
Extended Text Interaction	16	89	49	40
Other	1	11	24	80
PARCC Number of Points	Count	Percentage	Count	Percentage
EBSR -2 points	2	11	98	80
PCR Reading -3 points	8	44	8	7
PCR Reading -4 points	8	44	16	13
Text Complexity	Count	Percentage	Count	Percentage
Low	6	33	42	34
Medium	10	56	67	55
High	2	11	13	11
Task Type	Count	Percentage	Count	Percentage
Literary Analysis Task	7	39	24	20
Narrative Writing Task	4	22	20	16
Research Simulation Task	6	33	30	25
Other	1	6	48	39
Passage Type	Count	Percentage	Count	Percentage
Informational	6	33	54	44
Literary	12	67	68	56

Table A.2 Characteristics of Items Flagged for Flagged for Effect Sizes for Mathematics Selected Forms and Grade Levels

	Flagged Effect Size Items		All items	
Technology Enhanced Item Type	Count	Percentage	Count	Percentage
ConstructedResponse	5	20	15	13
ConstructedResponse:ConstructedResponse	4	16	5	4
ConstructedResponse:ConstructedResponse:ConstructedResponse:ConstructedResponse	1	4	2	2
FillInTheBlankSpecificCharacterSet	2	8	8	7
FillInTheBlankSpecificCharacterSet:FillInTheBlankSpecificCharacterSet	2	8	4	4
FillInTheBlankSpecificCharacterSet:FillInTheBlankSpecificCharacterSet:MultipleResponse:MultipleResponse	1	4	1	1
FillInTheBlankSpecificCharacterSet:FillInTheBlankSpecificCharacterSet:SelectedResponse:SelectedResponse	1	4	1	1
MultipleResponse	1	4	7	6
MultipleSelect	1	4	15	13
SelectedResponse	5	20	40	35
SelectedResponse:SelectedResponse	1	4	9	8
SelectedResponse:SelectedResponse:SelectedResponse:SelectedResponse	1	4	1	1
Other Technology Enhanced Item Type			5	4
Interaction Type	Count	Percentage	Count	Percentage
Choice Interaction	7	28	61	54
Extended Text Interaction	5	20	15	13
Text Entry Interaction	2	8	8	7
Other	11	44	29	26
Cognitive Complexity	Count	Percentage	Count	Percentage
Low	9	36	51	45
Medium	9	36	46	41
High	7	28	16	14
PARCC Number of Points	Count	Percentage	Count	Percentage
Type 1 - 1 point	9	36	69	61
Type 1 - 2 points	3	12	15	13
Type 1 - 4 points	3	12	5	4

	Flagged Effect Size Items		All items	
Type 2 - 3 points	1	4	4	4
Type 2 - 4 points	1	4	5	4
Type 3 - 3 points	4	16	9	8
Type 3 - 6 points	4	16	6	5
Response Type	Count	Percentage	Count	Percentage
Judgment	20	80	74	65
Multiple Choice	5	20	39	35

Appendix B

Table B.1 Summary of Mantel-Haenszel/SMD DIF Results for ELA/L Assessments by Grade-Level

Grade	DIF Category	Mantel-Haenszel/SMD	
		Total Number of Common Items	Percentage
3	A	46	88%
	B-	2	4%
	B+	4	8%
	C-	0	0%
	C+	0	0%
	Total	52	
4	A	50	88%
	B-	1	2%
	B+	5	9%
	C-	0	0%
	C+	1	2%
	Total	57	
5	A	61	98%
	B-	1	2%
	B+	0	0%
	C-	0	0%
	C+	0	0%
	Total	62	
6	A	55	98%
	B-	1	2%
	B+	0	0%
	C-	0	0%
	C+	0	0%
	Total	56	
7	A	74	93%
	B-	0	0%
	B+	6	8%
	C-	0	0%
	C+	0	0%
	Total	80	

Grade	DIF Category	Mantel-Haenszel/SMD	
		Total Number of Common Items	Percentage
8	A	78	95%
	B-	0	0%
	B+	3	4%
	C-	0	0%
	C+	1	1%
	Total	82	
9	A	72	90%
	B-	1	1%
	B+	5	6%
	C-	2	3%
	C+	0	0%
	Total	80	
10	A	63	98%
	B-	0	0%
	B+	1	2%
	C-	0	0%
	C+	0	0%
	Total	64	
11	A	77	100%
	B-	0	0%
	B+	0	0%
	C-	0	0%
	C+	0	0%
	Total	77	

Table B.2 Summary of Mantel-Haenszel/SMD DIF Results for Mathematics Assessments by Grade-Level

Grade	DIF Category	Mantel-Haenszel/SMD	
		Total Number of Common Items	Percentage
3	A	94	85%
	B-	9	8%
	B+	4	4%
	C-	4	4%
	C+	0	0%
	Total	111	
4	A	91	95%
	B-	1	1%
	B+	1	1%
	C-	3	3%
	C+	0	0%
	Total	96	
5	A	72	96%
	B-	1	1%
	B+	1	1%
	C-	0	0%
	C+	1	1%
	Total	75	
6	A	78	95%
	B-	0	0%
	B+	0	0%
	C-	3	4%
	C+	1	1%
	Total	82	
7	A	87	93%
	B-	2	2%
	B+	1	1%
	C-	3	3%
	C+	1	1%
	Total	94	
8	A	64	88%
	B-	6	8%
	B+	2	3%
	C-	1	1%
	C+	0	0%
	Total	73	
Algebra I	A	88	88%
	B-	7	7%
	B+	3	3%

Grade	DIF Category	Mantel-Haenszel/SMD	
		Total Number of Common Items	Percentage
	C-	2	2%
	C+	0	0%
	Total	100	
Geometry	A	92	79%
	B-	1	1%
	B+	14	12%
	C-	2	2%
	C+	7	6%
	Total	116	
Algebra II	A	59	75%
	B-	4	5%
	B+	4	5%
	C-	7	9%
	C+	5	6%
	Total	79	
Integrated Mathematics I	A	31	82%
	B-	3	8%
	B+	1	3%
	C-	3	8%
	C+	0	0%
	Total	38	
Integrated Mathematics II	A	38	78%
	B-	0	0%
	B+	7	14%
	C-	0	0%
	C+	4	8%
	Total	49	
Integrated Mathematics III	A	25	68%
	B-	6	16%
	B+	1	3%
	C-	1	3%
	C+	4	11%
	Total	37	

Appendix C

Table C.1 Characteristics of Items Flagged for C-Level DIF for ELA/L

	Flagged B-DIF Items		Flagged C-DIF		All Items	
Cognitive Complexity	Count	Percentage	Count	Percentage	Count	Percentage
Low	6	20	2	50	199	33
Medium	12	40	2	50	300	49
High	12	40			111	18
Response Type	Count	Percentage	Count	Percentage	Count	Percentage
Multiple:Multiple	1	3			23	4
Multiple:Single					9	1
Single:Multiple	1	3	1	25	94	15
Single:Multiple:Single					3	0
Single: Single:	5	17	2	50	352	58
Single: Single: Single	2	7			36	6
Single: Single: Single:Single	21	70	1	25	93	15
Interaction Type	Count	Percentage	Count	Percentage	Count	Percentage
Choice Interaction	5	17	1	25	254	42
Extended Text Interaction	23	77	1	25	123	20
Other	2	7	2	25	233	38
PARCC Number of Points	Count	Percentage	Count	Percentage	Count	Percentage
EBSR -2 points	7	23	3	75	487	80
PCR Reading -3 points	9	30	1	25	46	8
PCR Reading -4 points	14	47			77	13
Text Complexity	Count	Percentage	Count	Percentage	Count	Percentage
Low	15	50	1	25	195	32
Medium	8	27	2	50	307	50
High	7	23	1	25	108	18
Task Type	Count	Percentage	Count	Percentage	Count	Percentage
Literary Analysis Task	18	60	1	25	136	22
Narrative Writing Task	2	7	1	25	87	14
Research Simulation Task	5	17	1	25	135	22
Other	5	17	1	25	252	41
Passage Type	Count	Percentage	Count	Percentage	Count	Percentage
Informational	8	27	1	25	287	47
Literary	22	73	3	75	323	53

Table C.2 Characteristics of Items Flagged for C-Level DIF for Mathematics

	Flagged B-DIF Items		Flagged C-DIF		All items	
Technology Enhanced Item Type	Count	Percentage	Count	Percentage	Count	Percentage
ConstructedResponse	6	8	7	13	103	11
ConstructedResponse:ConstructedResponse	6	8	7	13	42	4
ConstructedResponse:ConstructedResponse:ConstructedResponse	1	1			11	1
ConstructedResponse:ConstructedResponse:ConstructedResponse:ConstructedResponse	1	1			3	0
FillInTheBlankSpecificCharacterSet	19	24	14	27	97	10
FillInTheBlankSpecificCharacterSet:FillInTheBlankSpecificCharacterSet	1	1			37	4
FillInTheBlankSpecificCharacterSet:MultipleResponse			1	2	1	0
MultipleResponse	7	9	5	10	66	7
MultipleResponse:FillInTheBlankSpecificCharacterSet			1	2	1	0
MultipleSelect	16	20	10	19	89	9
MultipleSelect:MultipleSelect	1	1	1	2	4	0
NotTechnologyEnhanced	1	1			3	0
SelectedResponse	18	23	6	12	363	38
SelectedResponse:ConstructedResponse	2	3			4	0
Other Technology Enhanced Item Type					126	13
Interaction Type	Count	Percentage	Count	Percentage	Count	Percentage
Choice Interaction	41	52	21	40	510	54
Extended Text Interaction	6	8	7	13	106	11
Text Entry Interaction	19	24	14	27	95	10
Other	13	16	10	19	239	25
Cognitive Complexity	Count	Percentage	Count	Percentage	Count	Percentage
Low	34	43	21	40	406	43
Medium	31	39	25	48	421	44
High	14	18	6	12	123	13
PARCC Number of Points	Count	Percentage	Count	Percentage	Count	Percentage
Type 1 - 1 point	60	76	35	67	606	64
Type 1 - 2 points	3	4	3	6	140	15

	Flagged B-DIF Items		Flagged C-DIF		All items	
Technology Enhanced Item Type	Count	Percentage	Count	Percentage	Count	Percentage
Type 1 - 4 points					23	2
Type 2 - 3 points	4	5	2	4	54	6
Type 2 - 4 points	4	5	4	8	44	5
Type 3 - 3 points	3	4	3	6	52	5
Type 3 - 6 points	5	6	5	10	31	3
Response Type	Count	Percentage	Count	Percentage	Count	Percentage
Judgment	61	77	46	88	594	63
Multiple Choice	18	23	6	12	356	37

Appendix D

Table D.1 Description of Item Response Types for PARCC ELA/L Assessments

Response Type	Description
Multiple: Multiple	Two-part Multiple Choice Multiple Select (MCMS) item
Multiple: Single	Two-part item that consists of a MCMS item followed by a Multiple Choice Single Select (MCSS) item
Single: Multiple	Two-part item that consists of a MCSS item followed by a MCMS item
Single: Multiple: Single	Three-part item that consists of a MCSS item followed by a MCMS items followed by a MCSS item
Single: Single	Two-part MCSS item
Single: Single: Single	Three-part MCSS item
Single: Single: Single: Single	Four-part MCSS item

Table D.2 Description of Technology Enhanced Item Types for PARCC Mathematics Assessments

Technology Enhanced Item Type	Description
ConstructedResponse	Constructed Response item
ConstructedResponse:ConstructedResponse	Two-part Constructed Response item
ConstructedResponse:ConstructedResponse:ConstructedResponse	Three-part Constructed Response item
ConstructedResponse:ConstructedResponse:ConstructedResponse: ConstructedResponse	Four-part Constructed Response item
FillInTheBlankSpecificCharacterSet	Fill-in-the-Blank item with Specific Character Set
FillInTheBlankSpecificCharacterSet:FillInTheBlankSpecificCharacter Set	Two-part Fill-in-the-Blank item with Specific Character Set
FillInTheBlankSpecificCharacterSet:MultipleResponse	Two-part item that consists of a Fill-in-the-Blank item with Specific Character Set item followed by a Multiple Response item
FillInTheBlankSpecificCharacterSet:FillInTheBlankSpecificCharacter Set:MultipleResponse:MultipleResponse	Four-part item that consists of two Fill-in-the-Blank items with Specific Character Sets followed by two Multiple Response items
FillInTheBlankSpecificCharacterSet:FillInTheBlankSpecificCharacter Set:SelectedResponse:SelectedResponse	Four-Part item that consists of two Fill-in-the-Blank items with Specific Character Sets followed by two Selected Response items
MultipleResponse	Multiple Response item
MultipleResponse:FillInTheBlankSpecificCharacterSet	Two-part item that consists of a Multiple Response item followed by a Fill-in-the-Blank item with Specific Character Set
MultipleSelect	Multiple Choice Multiple Select item
MultipleSelect:MultipleSelect	Two-part Multiple Choice Multiple Select item
NotTechnologyEnhanced	Not Technology Enhanced item
SelectedResponse	Selected Response item
SelectedResponse:ConstructedResponse	Two-part item that consists of a Selected Response item followed by a Constructed Response item
SelectedResponse:SelectedResponse	Three-part Selected Response item

Table D.3 Description of Item Interaction Types for PARCC Assessments

Interaction Type	Description
Choice Interaction	A set of choices is presented to the test taker. The test taker's task is to select one or more of the choices, up to a maximum number of choices.
Extended Text Interaction	Allows the test taker to enter a large block of text.
Text Entry Interaction	Allows the test taker to a simple piece of text.

Table D.4 Description of PARCC Number of Points

PARCC Number of Points	Description
EBSR -2 points	Two point Evidence Based Selected Response item
PCR Reading -3 points	Three point Prose Constructed Response item
PCR Reading -4 points	Four point Prose Constructed Response item
Type 1 - 1 point	One point task assessing concepts, skills, and procedures
Type 1 - 2 points	Two point task assessing concepts, skills, and procedures
Type 1 - 3 points	Three point task assessing concepts, skills, and procedures
Type 1 - 4 points	Four point task assessing concepts, skills, and procedures
Type 2 - 3 points	Three point task assessing expressing mathematical reasoning
Type 2 - 4 points	Four point task assessing expressing mathematical reasoning
Type 3 - 3 points	Three point task assessing modeling/applications
Type 3 - 4 points	Four point task assessing modeling/applications
Type 3 - 6 points	Six point task assessing modeling/applications