Evaluation of the impact of equating approach on the parameter and student score stability using pre- and post-equated designs in the post-pandemic environment

Joanna Tomkowicz, Dong-In Kim, and Ping Wan
Data Recognition Corporation

Paper presented at the annual meeting of
the National Council on Measurement in Education,
San Diego

April 23, 2022

Abstract

In this study we evaluated the stability of item parameters and student scores, using the pre-equated (pre-pandemic) parameters from Spring 2019 and post-equated (post-pandemic) parameters from Spring 2021 in two calibration and equating designs related to item parameter treatment: re-estimating all anchor parameters (Design 1) and holding the *c*-parameter fixed for all multiple-choice items (Design 2). The Spring 2021 English Language Arts and Mathematics grades 3 through 8 operational test data from a large-scale testing program were used for this study. It was found that re-estimating item parameters in both Design 1 and Design 2 had little effect on the calibration and equating results, indicating acceptable stability of item parameters re-estimated in the post-pandemic environment. There was little or no difference in the mean scale scores and percentages of students classified in the different performance levels when students were scored using the pre-equated versus post-equated item parameters in Spring 2021. These differences in performance were even smaller when scores obtained in post-equated Design 1 and post-equated Design 2 were compared. When Spring 2021 student performance was compared to the Spring 2019 student performance, a decrease in performance was observed in each grade and content area that could be attributed to the effect of the pandemic and disrupted learning in the 2020–21 school year.

Introduction and Study Purpose

Given school closures, large-scale assessment cancellations, and disruptions to instruction and student learning due to the Covid-19 pandemic in the 2020–21 school year, the use of pre-equated parameters in the scoring of students who participated in Spring 2021 testing was almost uniformly recommended by technical advisory committees and the leading experts in educational measurement (CCSSO, 2020). This recommendation resulted from concerns about the unknown effects of the pandemic on student learning in the 2020–21 school year and on student performance on the Spring 2021 assessments.

This study investigates the stability of item parameters and student scores using the pre- and post-equated (post-pandemic) design and three sets of item parameters: pre-equated parameters obtained in the Spring 2019 administration; re-estimated (post-equated) parameters for all items; and re-estimated (post-equated) item difficulty and discrimination parameters while holding the *c*-parameter for multiple-choice items fixed to their values from Spring 2019. This study contributes to a better understanding of the impact of choosing different approaches to the treatment of item parameters on student scores in the environment of disrupted learning.

Data and Method

Spring 2021 English Language Arts (ELA) and Mathematics grades 3 through 8 operational test data from a large-scale testing program were used in this study. All assessments included mixed item types (multiple-choice, multi-select, technology-enhanced, evidence-based selected-response, and short-answer) and were administered under standardized conditions in Spring 2021.

ELA operational test forms were reused from the Spring 2019 administration with modifications. The modifications to the test included the removal of an essay-like text-dependent analysis (TDA) item, which was replaced with two or three standalone autoscored items measuring the same Text Types and Purposes sub-domain in each grade. The Text Types and Purposes sub-domain is designed to assess a student's ability to identify, analyze, and/or produce quality writing. Each standard addresses a different type of writing: opinion/argumentative, informative/explanatory, or narrative. The sub-standards within each writing type target the specific features of quality writing, such as introductions, conclusions, transitions, and support (for the first two writing types) and narrative techniques, plot development, and characterization (for narrative). This change was implemented in order to reduce testing time in Spring 2021 and resulted in the ELA tests being approximately 45 minutes shorter than the Spring 2019 assessments. The form modifications did not negatively affect the test blueprint. Common items between the Spring 2019 and 2021 ELA administrations still constituted more than 90% of the assessment in each grade. Mathematics operational test forms were reused intact from the Spring 2019 administration.

For reporting purposes, students were scored using the item pattern method and pre-equated parameters for all items. For the purpose of this study, students were scored using the same scoring method and both pre- and post-equated sets of item parameters. The scale score statistics and percentages of students in different performance levels were computed and compared across the equating options. The stability of pre-equated scores was evaluated in the

post-equating verification using the two approaches to the item treatment that are outlined in the introduction section. Because the Spring 2021 forms were reused or reused with small modifications from the Spring 2019 administration, all test items were included in data calibration and all test items served as anchor items in post-equating.

Calibration samples

Student participation in the Spring 2021 administration was approximately 85% in grades 3 through 8. The Spring 2021 tested population was compared to the Spring 2019 tested population in regard to student gender, ethnicity, socioeconomic status, English language proficiency (ELP) status, disability status, and district locale as indicated in the National Center for Education Statistics database (https://nces.ed.gov/ccd/districtsearch).

The Spring 2021 tested population was found to be different from the Spring 2019 tested population regarding several demographic variables. Specifically, Black students, socioeconomically disadvantaged students, and students from districts designated as city districts were underrepresented in the Spring 2021 test participants, while White and not socioeconomically disadvantaged students were overrepresented. To align the student demographic characteristics between the Spring 2021 and Spring 2019 test administrations, sampling of the Spring 2021 data was performed using the propensity score matching (PSM) method (Rosenbaum & Rubin, 1983). The covariates in the PSM model included student gender, ethnicity, socioeconomic status, language proficiency status, disability status, and district locale.

In the first step of this procedure, representative samples of approximately 50% of test takers were selected from the Spring 2019 data using stratified random sampling for each grade. This sample became a target for selecting matching students from the Spring 2021 test takers. Second, comparable groups of students for data calibration and equating were created for each grade and content area using the Spring 2021 data. PSM allows modeling the conditional probability of assignment to a condition given a set of covariates. For the current study, we modeled the probability of a student taking the assessment in Spring 2019 based on a set of covariates. Spring 2021 to Spring 2019 student matching was performed using the "nearest neighbor" algorithm. The propensity score statistics for all grades are presented in Table 1 and provide evidence of acceptable student match between the two administrations. The 2021 calibration samples consisted of approximately 60% of tested students who had complete operational test data.

For illustration purposes, the characteristics of the Spring 2019 tested populations, the Spring 2021 tested populations, and the Spring 2021 calibration samples are presented in Table 2 for grade 3 (an elementary school grade) and in Table 3 for grade 7 (a middle school grade). As shown in these tables, the Spring 2021 calibration samples were comparable to the Spring 2019 state student population for the same grade. The differences between the demographic characteristics of the Spring 2021 calibration samples and all test takers in Spring 2019 were no larger than 0.7% for any demographic category (see the last column in Table 2 and Table 3). The same was found to be true for remaining grades and these results are not presented in this paper. Note that the data presented in Tables 2 and 3 are for the ELA samples. The characteristics of Mathematics test takers included in the calibration samples were very similar to those of ELA and are not presented in this paper. In all cases, the differences between the number of students

who took ELA and Mathematics assessments within the same grade, content area, and administration year were a half percent or less. These differences had no bearing on the overall tested population or calibration sample characteristics.

Calibration and Equating Designs

The Spring 2021 calibration sample data were used for data calibration and post-equating. All data were calibrated using the three-parameter (3PL) model for multiple-choice (MC) items and the two-parameter partial-credit (2PPC) model for all non-MC items (Lord & Novick, 1968; Yen & Fitzpatrick, 2006). Under the 3PL model, the probability that a student with the trait or scale score θ will respond correctly to MC item *j* is

$$P_j(\theta) = c_j + (1 - c_j)/[1 + \exp(-1.7a_j(\theta - b_j))].$$

In the equation, a_j is the item discrimination, b_j is the item difficulty, and c_j is the probability of a correct response by a very low-ability student. Under the 2PPC model, the probability that a student with the trait or scale score θ will respond in category *k* to partial-credit item *j* is

$$P_{jk}(\theta) = \exp(z_{jk}) / \sum_{i=1}^{m_j} \exp(z_{ji}),$$

where $z_{jk} = (k-1)f_j - \sum_{i=0}^{k-1} g_{ji}$, and $g_{j0} = 0$ for all *j*.

The resulting parameters estimated in the 3PL and 2PPC models are initially in two different metrics. The discrimination and location (difficulty) parameters for the MC items are in the traditional 3PL metric and are labeled *a* and *b*, respectively. In the 2PPC model, *f* (*alpha*) and *g* (*gamma*) are analogous to *a* and *b*, where *alpha* is the discrimination parameter and *gamma* over *alpha* (*g*/*f*) is the location where adjacent trace lines cross on the ability scale. Because of the different metrics used, the 3PL parameters *a* and *b* are not directly comparable to the 2PPC parameters *f* and *g*; however, they can be converted to a common metric. The two metrics are related by $b = g/f$ and $a = f / 1.7$ (Burket, 2002). As a result of this procedure, the MC and non-MC items are placed on the same scale. Note that for the 2PPC model, there are $m_j - 1$ (where m_j is a score level *j*) independent *g*'s and one *f*, for a total of m_j independent parameters estimated for each item, while there is one *a* and one *b* per item in the 3PL model.

In this study, the calibrations were conducted separately for each grade level and content area using the marginal maximum-likelihood procedures implemented with the expected maximum algorithm (Bock & Aitkin, 1981; Thissen, 1982). Two data calibration designs related to the treatment of test items were implemented: Design 1 in which all item parameters were re-estimated for all items and Design 2 in which difficulty and discrimination parameters were re-estimated for all items and the guessing parameters for MC items were fixed to their Spring 2019 (or most recent) values.

In the process of item calibration, the number of estimation cycles was set to 99 with the convergence criterion of 0.001 for all content areas. The maximum value of *a*- and *alpha*-parameters, reflecting item discrimination for MC and non-MC items, was set to 5.0, and the range for *b*- and *gamma*-parameters, reflecting item difficulty for MC and non-MC items, was set between -7.5 and 7.5 in both calibration designs. The maximum value of a *c*-parameter (the probability of guessing a correct response for an MC item) was set to 0.50 in calibration Design 1, where all item parameters were re-estimated.

Following data calibration, the test equating was performed using the Stocking & Lord (1983) equating procedure. A common-item design was used for post-equating the Spring 2021 assessments to the established ELA and Mathematics scales. All items on the test were used as anchor items. The pre-equated or anchor parameters were treated as "base" parameters. The stability of these anchor item parameters in both post-equating designs was evaluated using two statistical methods: (a) iterative linking (Candell & Drasgow, 1988) using Stocking & Lord's (1983) Test Characteristic Curve method and (b) computing differences between the item-ability regression curves.

The Stocking & Lord TCC method minimizes the mean squared difference between the two TCCs; one curve is based on estimates from the previous calibration (pre-pandemic), and the other curve is based on transformed estimates from the current calibration (post-pandemic). Differential item functioning for individual items was evaluated by examining pre-pandemic (input) and post-pandemic (estimated) item parameters. Items with an absolute difference of parameters greater than two times the root mean square deviation were flagged for differential functioning between the two administrations.

The IRT item-ability regression curve method was used to evaluate differences between the item-ability regression curves of the anchor items in Spring 2019 and Spring 2021. For this evaluation, the following measures were used: (1) unweighted mean signed difference in estimated probability; (2) unweighted mean absolute (unsigned) difference in estimated probability; (3) unweighted root mean squared difference; (4) weighted mean signed difference in estimated probability; (5) weighted mean absolute (unsigned) difference in estimated probability; (6) weighted root mean squared difference; and (7) the maximum absolute difference.

Both unweighted and weighted versions of the first three measures were calculated. Unweighted differences give equal weight to differences across the ability spectrum. Weighted differences assign weights according to the number of test takers that are impacted (that is, the frequency distribution of estimated student abilities during the calibration).

For the first six measures listed above, differences greater than +/-0.10 were considered large and differences between +/-0.07 and +/-0.10 were considered moderate. For the maximum absolute difference, large differences were those greater than +/-0.15 and moderate differences were differences between +/-0.125 and +/-0.15. Items were flagged if the item-ability regression curve measures met or exceeded the threshold for moderate differences.

Calibration Software

Calibration and equating of the data were performed using PARDUX software (Burket, 2002). PARDUX is designed to produce a single scale by jointly analyzing data resulting from students' responses to both MC items and CR items for assessments that include both item types. In PARDUX, items are calibrated based on IRT, using the 3PL model (Lord & Novick, 1968) for MC items and the 2PPC model (Yen, 1993) for non-MC items. PARDUX software has shown to produce parameter and ability estimates that were as precise as those estimated by PARSCALE (Muraki & Bock, 1991) and MULTILOG (Thissen, 1990) programs, which are widely known and used IRT programs (Fitzpatrick, 1991; Fitzpatrick and Julian, 1996).

Student Scoring

To evaluate the comparability of student scores estimated using pre-equated versus post-equated item parameters, the item pattern scoring method and the following three sets of item parameters were used to score students who took ELA and Mathematics tests in Spring 2021: pre-equated parameters (also used in score reporting), item parameters obtained in the calibration and equating Design 1 (in which all item parameters were re-estimated), and item parameters obtained in the calibration and equating Design 2 (in which *c*-parameters for MC items were held fixed to their pre-pandemic values).

In addition, the Spring 2021 student performance (for students included in the calibration samples) was compared to the Spring 2019 student performance (all test-takers) to assess changes in performance that could be attributed to disrupted learning in the 2020–21 school year. Because the Spring 2021 calibration samples were comparable with regard to student demographic characteristics to the Spring 2019 tested populations in each grade, the performance of students included in the calibration samples could be compared to that of the total tested population in Spring 2019. This comparison provides some insight into what the state assessment results might look like if the Spring 2021 tested population were comparable to the 2019 tested population. (Note, however, that the student performance results for students included in the calibration sample do not necessarily reflect the performance of all students who participated in the Spring 2021 assessment.)

Recall that while Spring 2021 Mathematics operational assessments were reused intact from Spring 2019, this was not the case for ELA operational assessments. The TDA item that was included in the Spring 2019 ELA assessment was removed and replaced by two or three autoscored items in the Spring 2021 assessment in each grade. To account for this modification and for the purpose of results validation, a second set of ELA scores based on items that were common between the two administrations was computed for students who took these assessments in Spring 2019 and Spring 2021. Again, pre-equated and post-equated item parameters were used for this purpose.

Results

The study results are presented in this section. Calibration results are discussed first, followed by equating results, and then summary of student performance.

Calibration Results

Selected calibration results are presented in Table 4 for ELA and Table 5 for Mathematics. Some similarities and differences were observed between the Design 1 and Design 2 results in each content area. First, a larger number of iterations were observed in calibration Design 1 (where all item parameters were re-estimated) compared to calibration Design 2 (with a fixed *c*-parameter) for all ELA grades and for all Mathematics grades except for grade 6. For ELA, the number of iterations ranged from 34 for grade 7 to the maximum allowed, 99, for grades 3, 4, 6, and 8 in calibration Design 1, and the number of iterations ranged from 10 to 30 across all grades in calibration Design 2 (see Table 4). For Mathematics, the number of iterations ranged from 15 for grades 3 and 4 to the maximum allowed, 99, for grades 5 and 8 in calibration Design 1, and the number of iterations ranged from 9 to 27 across all grades in calibration Design 2 (see Table 5).

As shown in Tables 4 and 5, the estimated *a*- and *b*-parameters for MC items as well as *alpha*- and *gamma*-parameters for non-MC items were within the prescribed parameter ranges in both calibration designs in all grades and both content areas. Furthermore, the ranges of the discrimination and difficulty parameters were found to be comparable between Design 1 and Design 2 within each grade level and content area.

Yen's Q_1 statistics (Yen 1981, 1984) were used to evaluate model-to-data fit for all test items in both calibration designs. The numbers of items flagged for poor fit are also shown in Table 4 for ELA and Table 5 for Mathematics. Between 1 and 3 items were flagged for poor fit in ELA grades 3, 4, 5, 6, and 8. Six items were flagged for poor fit in ELA grade 7. No items were flagged for poor fit in Mathematics grades 4 and 6. Between 1 and 3 items were flagged for poor fit in the remaining Mathematics grades. The same items were flagged in Design 1 and Design 2 within each grade and content area, suggesting that re-estimating *c*-parameters in calibration Design 1 versus holding *c*-parameters fixed to their Spring 2019 values in Design 2 had no impact on the item fit for MC items. In fact, of all items flagged for poor fit, only three items (one in ELA grade 3 and two in Mathematics grade 8) were MC items. All other poor-fitting items were non-MC items. Inspection of the observed-to-predicted item characteristic curve for each flagged item revealed that these items had empirical (observed) information that differed from the model in the lower-ability range, where there are fewer students to provide information at the tail end of the distribution. Items that only show poor fit at the tail ends of the distribution provide stable information about the majority of the students—those in the middle range of the distribution. Overall, the number of items flagged for poor fit in both the ELA and Mathematics assessments was small and given the location of the misfit on the ability scale, the poor fitting items were of little concern.

In addition, putting the number of items flagged for poor fit in Spring 2021 in perspective, a majority of the items flagged for poor fit in Spring 2021 calibrations were also flagged for poor fit in the Spring 2019 test administration. Thirteen out of sixteen ELA items flagged for poor fit in Spring 2021 were also flagged for poor fit in Spring 2019 data calibration. Five out of eight Mathematics items flagged for poor fit in Spring 2021 were also flagged for poor fit in Spring 2019. Additional items were flagged for poor fit in Spring 2019 but not in Spring 2021 (one item in ELA grade 3, two items in ELA grade 5, one item in Mathematics grade 3, and one item in Mathematics grade 7).

Equating and Anchor Item Evaluation Results

Stocking and Lord equating summary and anchor evaluation results using the TCC method are presented in Table 6 for ELA and in Table 7 for Mathematics. These tables summarize the following information for each content area and grade: number of anchors, number of iterations, equating constants (*A* and *B*), quadratic loss function (*F*), correlation between the *a*-parameter input and estimates, correlation between the *b*-parameter input and estimates, correlation between the *c*-parameter input and estimates in Design 1, numbers of MC outlier items as indicated by the root mean square deviation method, as well as the correlation between the *alpha*-parameter input and estimates and the correlation between the *gamma*-parameter input and estimates for non-MC items.

The overall alignment of the anchor TCCs was very good for all grades in both content areas in Design 1 and Design 2. For illustration purposes, TCCs for grade 3 (an elementary school grade) and grade 7 (a middle school grade) are presented in this paper. Figures 1 and 2 show the TCC alignment of the anchor set before and after equating for ELA grades 3 and 7, respectively. Figures 3 and 4 show the TCC alignment of the anchor set before and after equating for Mathematics grades 3 and 7, respectively. In these figures, the input anchor set TCC (before equating) is indicated by the dashed red line and the new anchor estimate TCC (after equating) is indicated by the solid blue line. No visual differences between the anchor and estimate TCC alignment were found when Design 1 and Design 2 TCCs were compared for the same grade and content area, suggesting that re-estimating the *c*-parameter versus holding the *c*-parameter fixed in equating had little bearing on the overall anchor and estimate TCC alignment. This observation was also true for all other grades, though the TCCs for the remaining grades are not presented in this paper. In support of these findings, the equating constants *A* and *B* were found to be comparable between Design 1 and Design 2. The difference between equating constant *A* in Design 1 and equating constant *A* in Design 2 was approximately 0.02 or less in each grade and content area except for ELA grade 8 where the difference was 0.05 (still small). Differences between equating constant *B* in Design 1 and equating constant *B* in Design 2 were approximately 0.1 or less across all grades and both content areas except for ELA grade 8 where the difference between the two equating constants was 0.2 (see Tables 6 and 7).

As presented in Table 6, the correlations between the *a*-parameter input and estimates ranged from 0.93 to 0.97 in Design 1 and from 0.98 to 0.99 in Design 2 for all ELA grades. The correlations between the *b*-parameter input and estimates ranged from 0.96 to 0.99 in Design 1 across all ELA grades. The correlation between the *b*-parameter input and estimates was 0.99 in Design 2 for all ELA grades. The overall number of MC anchor items flagged as either *a*- or *b*-parameter outliers was small and ranged from one to three items, depending on the grade level and content area.

Interestingly, there was no consistent pattern of specific items being flagged as *a*- or *b*-parameter outliers in Design 1 and Design 2. In some cases, the same items were flagged in both equating designs (for example, anchor item 2 was flagged as an *a*-parameter outlier and anchor item 7 was flagged as a *b*-parameter outlier in ELA grade 5). In other cases, different items were flagged in Design 1 and Design 2. For example, in ELA grade 6, anchor item 6 in Design 1 and anchor item 27 in Design 2 were flagged as *a*-parameter outliers; also in ELA grade 6, anchor item 2 in Design 1 and anchor items 1 and 6 in Design 2 were flagged as *b*-

parameter outliers. These differences in specific items being flagged between Design 1 and Design 2 result from different treatment of the *c*-parameter in each equating design.

The correlations between the *c*-parameter input and estimates ranged from 0.73 to 0.89 in Design 1 across all ELA grades. Between one and three MC anchor items were flagged as *c*-parameter outliers in each ELA grade.

As expected, the correlations between the *alpha*-parameter input and estimates and *gamma*-parameter input and estimates were not affected by the treatment of the *c*-parameter for MC items. These correlations ranged from 0.98 to 0.99 for the *alpha*-parameter and from 0.98 to 1.00 for the *gamma*-parameter across all ELA grades and were the same in Design 1 and Design 2 within each grade for a given parameter (see Table 6).

As shown in Table 7, for Mathematics, the correlations between the *a*-parameter input and estimates ranged from 0.92 to 0.97 in Design 1 (except in grade 7 where the correlation was 0.89) and from 0.93 to 0.98 in Design 2. The correlations between the *b*-parameter input and estimates ranged from 0.97 to 1.00 in Design 1 across all Mathematics grades. The correlation between the *b*-parameter input and estimates was 0.99 in Design 2 for all Mathematics grades. The overall number of MC anchor items flagged as either *a*- or *b*-parameter outliers was two or three items in each grade and equating design except in Design 2 for grade 7, where four items were flagged (two items were flagged as *a*-parameter outliers and two different items were flagged as *b*-parameter outliers). Similar to ELA, no specific pattern in anchor item flagging across Design 1 and Design 2 was observed for Mathematics, indicating that different treatment of the *c*-parameter in equating leads to different MC items being flagged as outliers.

The correlations between the *c*-parameter input and estimates ranged from 0.84 to 0.98 in Design 1 across all Mathematics grades. One or two MC anchor items were flagged as *c*-parameter outliers across all Mathematics grades.

The correlations between *alpha*-parameter input and estimates were at least 0.97 or higher across all Mathematics grades except for grade 5 where correlation was 0.91. The correlations between the *gamma*-parameter input and estimates ranged from 0.98 to 1.00 across all Mathematics grades (see Table 7).

Again, for illustration purposes, the input (pre-equated) and estimate parameter values from Design 1 and Design 2 are presented for grades 3 and 7 ELA and Mathematics. The ELA results are presented in Figures 5 and 6 for grades 3 and 7, respectively. The Mathematics parameter values are shown in Figures 7 and 8 for grades 3 and 7, respectively. As can be observed in Figures 5 through 8, the majority of *a*- or *alpha*-parameters and *b*- or *gamma*-parameter values show little difference between the pre-equated design, post-equated Design 1, and post-equated Design 2. The exceptions were items noted as outliers (see Table 6 for ELA and Table 7 for Mathematics). Larger discrepancies were found between the pre-equated and post-equated (Design 1) *c*-parameter values of some items. However, as demonstrated by the TCC curve alignment and equating results reported in Tables 6 and 7, these discrepancies did not appear to have an effect on the overall quality of equating. Similar patterns of item parameter differences across equating designs were found for all other grades.

Very few items were flagged using the item-ability regression anchor evaluation method. The one item that was flagged in equating Design 1 in ELA grade 7 and the one item that was flagged in equating Design 2 in ELA grade 7 were one and the same (see Table 8). One item was flagged in equating Design 2 in Mathematics grade 4, and two items were flagged in equating Design 1 in Mathematics grade 7 (see Table 9).

Given that some performance change on test items was expected due to the circumstances related to the Covid-19 pandemic, anchors were not considered for removal from the anchor set solely due to the difference in item parameters between administrations. All flagged anchor items were reviewed, and it was confirmed that no changes to item content, format, or scoring rules occurred between the Spring 2019 and Spring 2021 test administrations. Because the post-equating results were intended to be used for evaluating item parameter stability rather than for making adjustments to the impact data, no anchors were excluded from post-equating in either equating approach.

Student Performance

While some differences were observed between the pre- and post-equated parameters and also between the post-equated parameters obtained in equating Design 1 and equating Design 2, perhaps the most important indicator of the item parameter stability across equating designs is student scores. Scale score summary statistics were computed using the pre-equated item parameters and post-equated item parameters obtained in Design 1 and Design 2 for students in the calibration samples. In addition, the percentages of students classified in different performance levels were computed based on test scores obtained in each equating design.

The ELA scale score summaries, computed using all items on the test, are presented in Table 10. The corresponding impact data (percentages of students in different performance levels) are presented in Table 11. Student performance data from Spring 2019 are also included in these tables to illustrate changes in student performance between Spring 2019 (pre-pandemic) and Spring 2021 (post-pandemic).

As shown in Table 10, a decline in performance as reflected by differences in mean scale scores between Spring 2019 and Spring 2021 was observed in each ELA grade. The mean scale score decreases ranged from about 2 scale score points in grade 8 to over 7 scale score points in grade 3, regardless of the equating design implemented in Spring 2021. On the other hand, the mean scale score differences between the pre-equated design and both post-equated designs in Spring 2021 were less than half a score point in each grade. The mean scale score differences between post-equated Design 1 and post-equated Design 2 were even smaller and ranged from 0.02 scale score points in grades 4 and 7 to 0.12 in grade 6.

The scale score standard deviations were comparable within approximately one scale score point between Spring 2019 and Spring 2021 (for all equating designs). The scale score standard deviations were comparable within half a scale score point across pre- and post-equated designs in Spring 2021. The median scores were comparable within one scale score point across pre- and post-equated designs within each ELA grade (refer to Table 10).

The pattern of differences in the percentages of students classified in different performance levels between Spring 2019 and Spring 2021, between the pre-equated and post-equated designs in Spring 2021, and between post-equated Design 1 and post-equated Design 2 in Spring 2021 followed the pattern of mean scale score differences for ELA. As seen in Table 11, a decrease in the percentage of students classified as *Proficient* or *Advanced* was observed between Spring 2019 and Spring 2021 in all ELA grades. These decreases ranged from close to 2% in grade 8 to over 6% in grade 3, regardless of the equating design implemented in Spring 2021. The differences in the percentages of students classified as *Proficient* or *Advanced* between the pre-equated and post-equated designs were approximately 0.3% or less across all ELA grades. The differences in the percentages of students classified as *Proficient* or *Advanced* between the post-equated Design 1 and post-equated Design 2 were approximately 0.1% or less across all ELA grades.

In addition, when the percentages of students in each of the four performance levels—*Below Basic*, *Basic*, *Proficient*, and *Advanced*— were considered, the differences in the percentages of students classified in any performance level were less than half a percent across all equating designs implemented in Spring 2021 in all ELA grades (see Table 11).

The scale score and performance level summary data presented in Tables 10 and 11 indicate that using pre-equated parameters versus item parameters recalibrated in the post-pandemic environment had little impact on the resulting student scores and performance level classification. Re-estimating all item parameters (Design 1) versus holding the *c*-parameters fixed (Design 2) had no practical effect on the assessment results for ELA.

As a side note and as mentioned in the "Data and Method" section of this paper, students who took ELA assessments in Spring 2019 and 2021 were rescored using their responses to items that were common between the two administrations. The common items constituted more than 90% of the ELA test in each grade. Students were rescored using the pre- and post-equated item parameters obtained in calibration and equating Design 1 and Design 2. The scale score means and standard deviations obtained from scores based on only common items were comparable, on average, within half a scale score point, with the scale score means and standard deviations obtained from scores based on full tests across all equating designs in all ELA grades. Similarly, the percentages of students in different performance levels were comparable, on average, within half a percent when students were classified into performance levels using the two sets of scores across all Spring 2021 equating designs in all grades. This finding suggests that the impact of the ELA test modifications on the ELA test scores was minimal. Because of the similarity of the results from the two sets of ELA test scores, the summaries of the test scores based on only the common items across the two test administrations are not presented in this paper.

The Mathematics scale score summaries are presented in Table 12. The corresponding impact data (percentages of students in different performance levels) are presented in Table 13. As for ELA, the Spring 2019 student performance data are also included in these tables for comparison.

As shown in Table 12, a decline in performance, reflected by differences in mean scale scores, between Spring 2019 and Spring 2021 was observed in each Mathematics grade. The mean scale score decreases ranged from approximately 7 to 8 scale score points in grades 4 and 7 to approximately 11 scale score points in grade 6. The mean scale score differences between the pre-equated design and post-equated Design 1 and Design 2 in Spring 2021 were less than half a score point in grades 3 through 5. The mean scale score differences between the pre-equated design and post-equated Design 2 were also less than half a score point in grades 6 and 7. Slightly larger differences were observed between the mean scale scores from the pre-equated design and post-equated Design 1 in grades 6 through 8 and between the pre-equated design and post-equated Design 2 in grade 8. These differences ranged from 0.66 scale score points (the difference between the pre-equated design and post-equated Design 1 in grade 6) to 1.42 scale score points (the difference between the pre-equated design and post-equated Design 1 in grade 7). These differences are still considered to be small. The mean scale score differences between equating Design 1 and equating Design 2 were less than one scale score point in each Mathematics grade.

The scale score standard deviations of the Spring 2019 scores were smaller by approximately 2 to 4 points compared to the scale score standard deviations of the Spring 2021 scores for grades 3 through 5, indicating larger score variability in Spring 2021 for these grades, regardless of the equating design implemented. The scale score standard deviations were comparable within approximately one scale score point between Spring 2019 and Spring 2021 for grades 6 through 8 (for all Spring 2021 equating designs). The scale score standard deviations were comparable within one scale score point across all Spring 2021 pre- and post-equated designs in all Mathematics grades (except in grade 7 where the difference between the pre-equated design and post-equated Design 1 was 1.59 points). The median scores were comparable within one scale score point in the Spring 2021 pre- and post-equated designs in each Mathematics grade in Spring 2021 (see Table 12).

As expected, the pattern of differences in the percentages of students classified in different performance levels between Spring 2019 and Spring 2021, between the pre-equated and post-equated designs in Spring 2021, and between Design 1 and Design 2 in Spring 2021 followed the pattern of mean scale score differences for Mathematics. As shown in Table 13, a decrease in the percentage of students classified as *Proficient* or *Advanced* was observed between Spring 2019 and Spring 2021 for all Mathematics grades, ranging from approximately 5% in grades 4 and 7 to about 8% in grade 6, regardless of the equating design implemented in Spring 2021. The differences in the percentages of students classified as *Proficient* or *Advanced* between the pre-equated and post-equated designs were approximately less than 0.3% across all grades except for grade 7 where the difference was about 0.6%. The differences in the percentages of students classified as *Proficient* or *Advanced* between post-equated Design 1 and post-equated Design 2 were less than 0.1% in all Mathematics grades.

In addition, when the percentages of students in each of the four proficiency levels—*Below Basic*, *Basic*, *Proficient*, and *Advanced*—were considered, the differences in the percentages of students classified in any performance level were less than half a percent between all equating designs in Mathematics grades 3 through 6 and less than one percent between all equating designs in Mathematics grades 7 and 8 in Spring 2021 (see Table 13).

The Mathematics scale score and performance level summary data presented in Tables 12 and 13 indicate that using pre-equated parameters versus item parameters recalibrated in the post-pandemic environment had little or no practical impact on the resulting student scores and performance level classification. In addition, re-estimating all item parameters (Design 1) versus holding the *c*-parameters fixed (Design 2) in the calibration had very little or no effect on the calibration and subsequent test results for Mathematics. These findings are consistent with the results obtained for ELA.

Summary

In summary, the pre-equated item parameters that were obtained after the Spring 2019 test administrations were used to score students who participated in the Spring 2021 assessments. This approach was recommended by many technical advisory committees and the leading experts in educational measurement (CCSSO, 2020) and was adopted out of an abundance of caution given the adverse effects of the Covid-19 pandemic on student learning in the 2020–21 school year.

The post-equating verification study was conducted to evaluate the parameter stability and the comparability of scale scores estimated using the pre- and post-equated parameters obtained in post-equated Design 1 (in which all parameters were re-estimated in calibration) and post-equated Design 2 (in which the *c*-parameters for MC items were held fixed to their prior values in calibration). The calibration results were satisfactory, and very few items were flagged for poor fit in Design 1 and Design 2 across all grade levels and both content areas. Re-estimating all item parameters versus holding the *c*-parameters fixed appeared to have little effect on the resulting item parameters. The equating results showed high correlations between the input and estimate parameters and very good alignment of input (anchor or pre-equated) and estimate TCCs. The number of items flagged using the TCC method and item-ability regression method was small in both Design 1 and Design 2. The calibration and equating results indicate overall acceptable stability of item parameters re-estimated in the post-pandemic environment.

A decrease in student performance between Spring 2019 and Spring 2021 was both expected and observed in each grade and content area. This change in performance, which was significantly larger than what is considered a typical year-to-year fluctuation of student scores, may serve as evidence of the effect of the Covid-19 pandemic on student achievement in the 2020–21 school year.

When only Spring 2021 student performance was considered, the differences found between the mean scale scores, scale score standard deviations, and percentages of students classified in the four performance levels were small and of no practical importance based on the use of the pre-equated versus post-equated item parameters. The differences were even smaller when the mean scale scores, scale score standard deviations, and percentages of students classified in the four performance levels were compared between post-equated Design 1 and post-equated Design 2 for both ELA and Mathematics.

In conclusion, while the use of pre-equated parameters in both ELA and Mathematics assessments was appropriate and justified given the uncertainty about the impact of the pandemic and disruptions to student learning on student performance on these assessments, post-equating

of the assessments would have resulted in comparable student scores and percentages of students in different performance levels. This finding provides much-needed evidence that post-equated parameters can be used effectively in the post-pandemic environment for future form building, equating, or student scoring, if such needs arise.

References

Burket, G. R. (2002). PARDUX [Computer program]. Unpublished.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46*, 443–459.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing bias in item response theory. *Applied Psychological Measurement*, *12*(3), 253–260.

Council of Chief State School Officers. (2020). Restart and recovery: Assessments in Spring 2021. https://www.nciea.org/sites/default/files/publications/Assessments-Spring-2021.pdf

Fitzpatrick, A. R. (1991). *Status report on the results of preliminary analyses of dichotomous and multi-level items using the PARMATE program*. Monterey, CA: CTB/McGraw-Hill.

Fitzpatrick, A. R., & Julian, M. W. (1996). *Two studies comparing the parameter estimates produced by PARDUX and PARSCALE*. Unpublished manuscript.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Muraki, E., & Bock, R. D. (1991). PARSCALE: Parameter scaling of rating data [Computer program]. Chicago, IL: Scientific Software, Inc.

National Center for Education Statistics. (2021). School District Characteristic 2018–2019 [Data file]. Retrieved from https://nces.ed.gov/ccd/districtsearch/

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 17*(1), 41–55.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210.

Thissen, D. (1982). Marginal maximum-likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47*(2), 175–186.

Thissen, D. (1990). MULTILOG: Multiple categorical item analysis and test scoring (Version 6) [Computer program]. Chicago, IL: Scientific Software, Inc.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245–262.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21*(2), 93–111.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213.

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

Table 1. Propensity Score Matching Results

Grade	Student Groups	Spring 2019 (Treatment Group)			Spring 2021 (Control Group)			Propensity Score Mean Difference Treatment - Control
		N-count	Mean Propensity Score	Propensity Score Std. Dev.	N-count	Mean Propensity Score	Propensity Score Std. Dev.	
3	All records (before matching)	30546	0.373	0.049	52720	0.364	0.044	0.009
	Matched records	30474	0.372	0.049	30474	0.372	0.048	0.000
4	All records (before matching)	31764	0.383	0.050	52501	0.373	0.044	0.009
	Matched records	31563	0.382	0.049	31563	0.382	0.049	0.000
5	All records (before matching)	32327	0.381	0.049	53793	0.372	0.044	0.009
	Matched records	32198	0.381	0.048	32198	0.380	0.048	0.000
6	All records (before matching)	32693	0.377	0.047	55198	0.369	0.041	0.008
	Matched records	32492	0.376	0.046	32492	0.376	0.046	0.000
7	All records (before matching)	31939	0.368	0.047	56013	0.360	0.042	0.008
	Matched records	31812	0.368	0.046	31812	0.368	0.046	0.000
8	All records (before matching)	31528	0.363	0.041	56411	0.356	0.037	0.006
	Matched records	31528	0.363	0.041	31528	0.363	0.041	0.000

Table 2. Spring 2019 Tested Population, Spring 2021 Tested Population, and Spring 2021 Calibration Sample Characteristics, Grade 3

Grade	Demo. Variable	Student Group	Total Spring 2019 N-count	Total Spring 2019 %	Total Spring 2021 N-count	Total Spring 2021 %	Difference % Total Spring 2021-Total Spring 2019	Calib. Sample Spring 2021 N-count	Calib. Sample Spring 2021 %	Difference % Calib. Sample Spring 2021-Total Spring 2019
3	All Students	All Students	61091	100.0	52930	100.0		30474	100.0	
	Gender	Female	29974	49.1	25886	48.9	-0.2	14891	48.9	-0.2
		Male	31117	50.9	27044	51.1	0.2	15583	51.1	0.2
	Race/ Ethnicity	American Indian	726	1.2	563	1.1	-0.1	297	1.0	-0.2
		Asian	2512	4.1	2237	4.2	0.0	1304	4.3	0.2
		African American	6565	10.7	4101	7.7	-3.0	3235	10.6	-0.1
		Hispanic	8295	13.6	6787	12.8	-0.8	4108	13.5	-0.1
		White	40204	65.8	36547	69.0	3.2	20141	66.1	0.3
		Two or More	2745	4.5	2695	5.1	0.6	1389	4.6	0.1
	Disability	No	53064	86.9	46009	86.9	0.1	26558	87.1	0.3
		Yes	8027	13.1	6921	13.1	-0.1	3916	12.9	-0.3
	Limited English Proficiency	No	55479	90.8	48597	91.8	1.0	27717	91.0	0.1
		Yes	5612	9.2	4333	8.2	-1.0	2757	9.0	-0.1
	Economically Disadvantaged	No	33672	55.1	32393	61.2	6.1	16870	55.4	0.2
		Yes	27419	44.9	20537	38.8	-6.1	13604	44.6	-0.2
	District Locale	Non-public	2989	4.9	3143	5.9	1.0	1479	4.9	0.0
		City	18707	30.6	12798	24.2	-6.4	9280	30.5	-0.2
		Suburban	16667	27.3	15705	29.7	2.4	8330	27.3	0.1
		Town	11901	19.5	10799	20.4	0.9	5969	19.6	0.1
		Rural	10827	17.7	10275	19.4	1.7	5416	17.8	0.0

Table 3. Spring 2019 Tested Population, Spring 2021 Tested Population, and Spring 2021 Calibration Sample Characteristics, Grade 7

Grade	Demo. Variable	Student Group	Total Spring 2019 N-count	Total Spring 2019 %	Total Spring 2021 N-count	Total Spring 2021 %	Difference % Total Spring 2021-Total Spring 2019	Calib. Sample Spring 2021 N-count	Calib. Sample Spring 2021 %	Difference % Calib. Sample Spring 2021-Total Spring 2019
7	All Students	All Students	63878	100.0	56295	100.0		31812	100.0	
	Gender	Female	31092	48.7	27448	48.8	0.1	15287	48.1	-0.6
		Male	32786	51.3	28847	51.2	-0.1	16525	51.9	0.6
	Race/ Ethnicity	American Indian	805	1.3	568	1.0	-0.3	357	1.1	-0.1
		Asian	2493	3.9	2129	3.7	-0.2	1313	4.1	0.2
		African American	6573	10.3	4251	7.6	-2.7	3153	9.9	-0.4
		Hispanic	8672	13.6	7199	12.8	-0.8	4349	13.7	0.1
		White	42845	67.1	39754	70.6	3.5	21399	67.3	0.2
		Two or More	2444	3.8	2394	4.3	0.4	1241	3.9	0.1
	Disability	No	56166	87.9	49987	88.8	0.9	28050	88.2	0.2
		Yes	7712	12.1	6308	11.2	-0.9	3762	11.8	-0.2
	Limited English Proficiency	No	60272	94.4	53350	94.8	0.4	29990	94.3	-0.1
		Yes	3606	5.6	2945	5.2	-0.4	1822	5.7	0.1
	Economically Disadvantaged	No	36985	57.9	35697	63.4	5.5	18438	58.0	0.1
		Yes	26893	42.1	20598	36.6	-5.5	13374	42.0	-0.1
	District Locale	Non-public	2993	4.7	3122	5.5	0.9	1484	4.7	0.0
		City	18131	28.4	12543	22.3	-6.1	8946	28.1	-0.3
		Suburban	18146	28.4	16785	29.8	1.4	9121	28.7	0.3
		Town	12796	20.0	12143	21.6	1.5	6385	20.1	0.0
		Rural	11812	18.5	11420	20.3	1.8	5876	18.5	0.0

Table 4. Spring 2021 Item Calibration Summary, English Language Arts

Grade	Design	# of All Items and Points	# of MC Items	# of non-MC Items	N-count	# of Iterations	Test Reliability	A/Alpha Parameter Range	B-Parameter Range	C-Parameter Range	Gamma 1 Range	Gamma 2 Range	# of Items with Poor Fit
3	1	38 (48)	27	11	30474	99	0.896	0.35703 to 1.5046	-0.7656 to 1.5238	0.0523 to 0.3177	-2.4928 to 1.1065	-0.6426 to 1.2973	2
3	2	38 (48)	27	11	30474	14	0.896	0.36399 to 1.53813	-0.7028 to 1.5786	0.0365 to 0.404	-2.3751 to 1.205	-0.5867 to 1.3732	2
4	1	41 (51)	28	13	31563	99	0.895	0.29103 to 1.27015	-1.0204 to 1.4211	0 to 0.2904	-2.8581 to 0.9367	-0.5671 to 1.6441	1
4	2	41 (51)	28	13	31563	15	0.895	0.30794 to 1.31095	-1.0069 to 1.4862	0.0385 to 0.2905	-2.8176 to 0.9653	-0.5257 to 1.6939	1
5	1	42 (51)	29	13	32198	43	0.900	0.27087 to 2.14909	-1.4168 to 1.1827	0.0654 to 0.3543	-2.8484 to -0.1042	-0.5644 to 0.6902	2
5	2	42 (51)	29	13	32198	16	0.900	0.27809 to 2.1976	-1.1899 to 1.3395	0.0607 to 0.3154	-2.571 to 0.003	-0.4712 to 0.7733	2
6	1	39 (51)	23	16	32492	99	0.885	0.33763 to 1.37922	-0.9212 to 1.5982	0.0861 to 0.2665	-2.4679 to 0.4528	-1.072 to 0.9184	2
6	2	39 (51)	23	16	32492	10	0.885	0.34253 to 1.39891	-0.9639 to 1.6324	0 to 0.2773	-2.4309 to 0.5154	-1.0266 to 0.9548	2
7	1	37 (51)	20	17	31812	34	0.886	0.35861 to 1.27284	-0.7728 to 1.1846	0.0521 to 0.2701	-2.3487 to 0.9102	-1.727 to 1.5121	6
7	2	37 (51)	20	17	31812	16	0.886	0.36257 to 1.29255	-0.5053 to 1.2338	0.0494 to 0.3084	-2.3073 to 0.9947	-1.6789 to 1.5355	6
8	1	40 (51)	29	11	31528	99	0.902	0.35722 to 1.49108	-1.5459 to 1.449	0.0337 to 0.3039	-2.5373 to 1.378	-1.5821 to 1.1806	3
8	2	40 (51)	29	11	31528	30	0.902	0.35842 to 1.57609	-1.1049 to 1.5202	0.0531 to 0.2961	-2.3894 to 1.4637	-1.4322 to 1.2635	3

Note: ELA non-MC items are worth 1 or 2 points.

Table 5. Spring 2021 Item Calibration Summary, Mathematics

Grade	Design	# of Items	# of MC Items	# of non-MC Items	N-count	# of Iterations	Test Reliability	A/Alpha Parameter Range	B-Parameter Range	C-Parameter Range	Gamma 1 Range	# of Items with Poor Fit
3	1	42	23	19	30329	15	0.931	0.46929 to 2.54028	-0.9287 to 1.3122	0.0875 to 0.371	-2.0558 to 2.0577	1
3	2	42	23	19	30329	9	0.931	0.50428 to 2.50658	-0.9107 to 1.4074	0.0576 to 0.3823	-2.0363 to 2.0772	1
4	1	46	32	14	31400	15	0.920	0.56618 to 2.36158	-1.5099 to 2.3258	0.0273 to 0.3277	-0.8881 to 2.5974	0
4	2	46	32	14	31400	12	0.920	0.46434 to 2.35331	-1.5058 to 2.9636	0.0255 to 0.3774	-0.8834 to 2.6014	0
5	1	46	27	19	31971	99	0.920	0.53775 to 1.91326	-0.9516 to 2.356	0.0638 to 0.483	-0.5783 to 2.2725	3
5	2	46	27	19	31971	17	0.920	0.54091 to 1.9132	-0.959 to 2.4807	0.0682 to 0.447	-0.5317 to 2.3163	3
6	1	46	31	15	32273	21	0.910	0.4191 to 2.42554	-1.2433 to 2.1737	0.0678 to 0.4244	-0.7458 to 3.1969	0
6	2	46	31	15	32273	23	0.910	0.43089 to 2.41744	-1.3101 to 2.2165	0.0807 to 0.3993	-0.7382 to 3.205	0
7	1	46	31	15	31604	31	0.903	0.34955 to 2.9215	-1.0362 to 2.5195	0.0576 to 0.3546	-0.9702 to 4.4851	2
7	2	46	31	15	31604	27	0.903	0.34903 to 2.94388	-0.7470 to 2.5020	0.0451 to 0.4222	-0.9501 to 4.5283	2
8	1	46	32	14	31296	99	0.910	0.31346 to 3.02199	-0.956 to 2.1158	0.0372 to 0.3603	-0.3238 to 2.6662	2
8	2	46	32	14	31296	17	0.910	0.42231 to 3.07093	-0.7525 to 2.1611	0.0306 to 0.3503	-0.1962 to 2.8986	2

Note: All Mathematics items are worth 1 point each.

Table 6. Equating and Anchor Evaluation Results Using the TCC Method, English Language Arts

Grade	Design	Equating Constants		# of Anchor Items	# of Iterations	*F* Value	MC Items						
							# of Items	A-Parameter		B-Parameter		C-Parameter	
		A	*B*					Corr.	# of Outliers*	Corr.	# of Outliers*	Corr.	# of Outliers*
3	1	0.9493	-1.3627	38	3	0.0887	27	0.93	2 (21, 37)	0.98	2 (6, 37)	0.88	3 (6, 21, 37)
3	2	0.9698	-1.4475	38	4	0.0595	27	0.98	2 (23, 30)	0.99	1 (4)		
4	1	1.0502	-0.7372	41	3	0.0941	28	0.96	1 (11)	0.97	2 (3, 31)	0.85	2 (6, 31)
4	2	1.0587	-0.7934	41	4	0.0526	28	0.98	2 (7, 11)	0.99	1 (30)		
5	1	0.9985	-0.3410	42	8	0.0740	29	0.97	1 (2)	0.97	1 (7)	0.85	1 (14)
5	2	1.0192	-0.4690	42	6	0.0186	29	0.99	2 (2, 7)	0.99	1 (7)		
6	1	1.0293	-0.1379	39	4	0.0425	23	0.97	1 (6)	0.96	1 (2)	0.73	1 (2)
6	2	1.0459	-0.1845	39	3	0.0936	23	0.99	1 (27)	0.99	2 (1, 6)		
7	1	1.1360	0.3586	37	5	0.1447	20	0.97	1 (6)	0.98	2 (4, 17)	0.87	2 (9, 17)
7	2	1.1461	0.2837	37	4	0.0678	20	0.98	1 (6)	0.99	0		
8	1	1.1760	0.5671	40	6	0.0724	29	0.97	0	0.99	1 (4)	0.89	3 (1, 10, 32)
8	2	1.2234	0.3691	40	4	0.0621	29	0.98	2 (12, 26)	0.99	1 (4)		

*Item numbers are provided in parentheses.

Table 6. Equating and Anchor Evaluation Results Using the TCC Method, English Language Arts (cont.)

Grade	Design	Non-MC Items					
		Alpha (*f*) Parameter		Gamma 1 Parameter		Gamma 2 Parameter	
		Corr.	# of Items	Corr.	# of Items	Corr.	# of Items
3	1	0.98	11	0.99	11	0.99	10
3	2	0.98	11	0.99	11	0.99	10
4	1	0.98	13	1.00	13	0.99	10
4	2	0.98	13	1.00	13	0.99	10
5	1	0.99	13	0.99	13	0.98	9
5	2	0.99	13	0.99	13	0.98	9
6	1	0.98	16	0.99	16	0.99	12
6	2	0.98	16	0.99	16	0.99	12
7	1	0.99	17	0.99	17	1.00	14
7	2	0.99	17	0.99	17	1.00	14
8	1	0.99	11	0.99	11	0.99	11
8	2	0.99	11	0.99	11	0.99	11

Table 7. Equating and Anchor Evaluation Results Using the TCC Method, Mathematics

Grade	Design	Equating Constants		# of Anchor Items	# of Iterations	*F* Value	MC Items						
							# of Items	A Parameter		B Parameter		C Parameter	
		A	*B*					Corr.	# of Outliers*	Corr.	# of Outliers*	Corr.	# of Outliers*
3	1	1.0088	-1.3801	42	6	0.0301	23	0.97	0	0.98	2 (4, 38)	0.90	1 (38)
3	2	0.9925	-1.3901	42	6	0.0611	23	0.98	1 (19)	0.99	2 (4, 19)		
4	1	0.9881	-0.8799	46	14	0.1167	32	0.92	2 (15, 31)	1.00	0	0.89	2 (1, 25)
4	2	0.9835	-0.8837	46	18	0.0273	32	0.93	2 (15, 42)	0.99	1 (42)		
5	1	0.9574	-0.3152	46	24	0.0792	27	0.96	1 (20)	0.99	1 (3)	0.98	1 (3)
5	2	0.9555	-0.3422	46	28	0.1013	27	0.98	1 (20)	0.99	1 (3)		
6	1	1.0565	-0.1981	46	27	0.1750	31	0.95	1 (28)	0.99	2 (8, 39)	0.95	2 (8, 12)
6	2	1.0510	-0.2001	46	26	0.0840	31	0.95	1 (32)	0.99	2 (12, 39)		
7	1	1.0338	0.2476	46	31	0.0575	31	0.89	2 (9, 22)	0.97	1 (12)	0.84	2 (12, 22)
7	2	1.0459	0.2288	46	23	0.1069	31	0.96	2 (9, 20)	0.99	2 (11, 12)		
8	1	1.0088	0.6716	46	31	0.0924	32	0.96	1 (2)	0.99	1 (38)	0.94	2 (4, 38)
8	2	1.0284	0.5898	46	33	0.1005	32	0.97	2 (1, 2)	0.99	1 (10)		

*Item numbers are provided in parentheses.

Table 7. Equating and Anchor Evaluation Results Using the TCC Method, Mathematics (cont.)

Grade	Design	Non-MC Items			
		Alpha (*f*) Parameter		Gamma 1 Parameter	
		Corr.	# of Items	Corr.	# of Items
3	1	0.97	19	0.99	19
3	2	0.97	19	0.99	19
4	1	0.98	14	0.99	14
4	2	0.98	14	0.99	14
5	1	0.91	19	0.98	19
5	2	0.91	19	0.98	19
6	1	0.98	15	0.99	15
6	2	0.98	15	0.99	15
7	1	0.99	15	0.99	15
7	2	0.99	15	1.00	15
8	1	0.99	14	0.99	14
8	2	0.99	14	0.99	14

Table 8. Anchor Evaluation Results Using the Item-Ability Regression Method, English Language Arts

Grade	Design	Item Type	Anchor Item Number	Unweighted Flags			Weighted Flags			Maximum Absolute Difference Flag	Total Number of IRT Flags
				Root Mean Squared Difference	Mean Absolute Difference	Mean Difference	Root Mean Square Difference	Mean Absolute Difference	Mean Difference		
7	1	Non-MC	25	—	—	—	Moderate	—	—	—	1
7	2	Non-MC	25	—	—	—	Moderate	—	—	—	1

Table 9. Anchor Evaluation Results Using the Item-Ability Regression Method, Mathematics

Grade	Design	Item Type	Anchor Item Number	Unweighted Flags			Weighted Flags			Maximum Absolute Difference Flag	Total Number of IRT Flags
				Root Mean Squared Difference	Mean Absolute Difference	Mean Difference	Root Mean Square Difference	Mean Absolute Difference	Mean Difference		
4	2	MC	42	Moderate	—	—	—	—	—	Large	2
7	1	MC	12	—	—	—	—	—	—	Moderate	1
7	1	MC	22	Moderate	—	—	—	—	—	Large	2

Table 10. Scale Score Summaries, English Language Arts

Grade	Design	Year	# of Items	# of Points	N-count	Scale Score Mean	Scale Score Std. Dev.	Median	Min. Score	Max. Score	Mean Difference		
											Spring 2021 - Spring 2019	Spring 2021 Pre-equated - Post-equated	Spring 2021 Design 1 - Design 2
3	Spring 2019	2019	37	53	61019	554.68	45.49	557	330	900			
3	Pre-equated	2021	38	48	30474	547.52	46.73	548	330	900	-7.15		
3	Post-equated Design 1	2021	38	48	30474	547.28	46.77	548	330	900	-7.39	0.24	
3	Post-equated Design 2	2021	38	48	30474	547.32	46.79	548	330	900	-7.35	0.20	-0.04
4	Spring 2019	2019	39	56	63444	582.11	50.98	583	340	930			
4	Pre-equated	2021	41	51	31563	575.58	51.35	577	340	930	-6.54		
4	Post-equated Design 1	2021	41	51	31563	575.93	51.28	577	340	930	-6.18	-0.36	
4	Post-equated Design 2	2021	41	51	31563	575.95	51.05	577	340	930	-6.16	-0.37	-0.02
5	Spring 2019	2019	40	56	64578	595.68	48.71	597	350	940			
5	Pre-equated	2021	42	51	32198	590.46	49.48	591	350	940	-5.21		
5	Post-equated Design 1	2021	42	51	32198	590.70	49.72	591	350	940	-4.98	-0.24	
5	Post-equated Design 2	2021	42	51	32198	590.64	49.40	591	350	940	-5.03	-0.18	0.05
6	Spring 2019	2019	37	56	65279	607.15	50.01	610	360	950			
6	Pre-equated	2021	39	51	32492	602.55	50.26	606	360	950	-4.60		
6	Post-equated Design 1	2021	39	51	32492	602.32	50.16	605	360	950	-4.83	0.23	
6	Post-equated Design 2	2021	39	51	32492	602.44	50.16	605	360	950	-4.71	0.11	-0.12
7	Spring 2019	2019	36	56	63767	627.84	54.74	631	370	960			
7	Pre-equated	2021	37	51	31812	622.87	55.45	625	370	960	-4.96		
7	Post-equated Design 1	2021	37	51	31812	623.33	55.59	626	370	960	-4.51	-0.46	
7	Post-equated Design 2	2021	37	51	31812	623.31	55.43	626	370	960	-4.53	-0.43	0.02
8	Spring 2019	2019	39	56	62914	629.30	59.61	633	380	970			
8	Pre-equated	2021	40	51	31528	627.29	58.89	630	380	970	-2.01		
8	Post-equated Design 1	2021	40	51	31528	627.05	58.73	630	380	970	-2.25	0.24	
8	Post-equated Design 2	2021	40	51	31528	627.09	58.80	630	380	970	-2.21	0.20	-0.04

Table 11. Performance Level Summary, English Language Arts

Grade	Design	Year	Percentage of Students in Performance Levels					Difference in Percentage of Students Classified as *Proficient* or *Advanced*		
			Below Basic	***Basic***	***Proficient***	***Advanced***	***Proficient* or *Advanced***	**Spring 2021 - Spring 2019**	**Spring 2021 Pre-equated - Post-equated**	**Spring 2021 Design 1 - Design 2**
3	Spring 2019	2019	23.22	38.04	33.25	5.49	38.74			
3	Pre-equated	2021	30.05	37.33	27.97	4.65	32.62	-6.12		
3	Post-equated Design 1	2021	29.97	37.35	28.22	4.46	32.68	-6.06	-0.06	
3	Post-equated Design 2	2021	30.06	37.34	28.09	4.52	32.60	-6.14	0.02	0.08
4	Spring 2019	2019	23.81	33.16	34.13	8.90	43.03			
4	Pre-equated	2021	28.99	32.32	31.57	7.12	38.69	-4.34		
4	Post-equated Design 1	2021	28.54	32.80	31.39	7.27	38.66	-4.37	0.03	
4	Post-equated Design 2	2021	28.60	32.66	31.58	7.16	38.74	-4.30	-0.04	-0.08
5	Spring 2019	2019	26.05	33.85	34.38	5.73	40.10			
5	Pre-equated	2021	30.01	34.43	30.76	4.79	35.55	-4.55		
5	Post-equated Design 1	2021	29.80	34.59	30.62	4.99	35.61	-4.49	-0.06	
5	Post-equated Design 2	2021	29.78	34.59	30.80	4.83	35.63	-4.47	-0.08	-0.02
6	Spring 2019	2019	23.47	35.50	31.93	9.10	41.03			
6	Pre-equated	2021	26.47	36.22	29.72	7.59	37.31	-3.72		
6	Post-equated Design 1	2021	26.68	36.33	29.45	7.54	37.00	-4.03	0.32	
6	Post-equated Design 2	2021	26.71	36.29	29.29	7.71	37.00	-4.03	0.31	-0.01
7	Spring 2019	2019	21.79	33.28	35.41	9.52	44.93			
7	Pre-equated	2021	24.94	33.81	32.92	8.33	41.25	-3.68		
7	Post-equated Design 1	2021	24.73	33.92	32.83	8.52	41.35	-3.58	-0.11	
7	Post-equated Design 2	2021	24.65	33.87	33.11	8.37	41.48	-3.45	-0.24	-0.13
8	Spring 2019	2019	25.80	37.09	28.86	8.25	37.11			
8	Pre-equated	2021	26.93	37.77	27.75	7.55	35.30	-1.81		
8	Post-equated Design 1	2021	27.33	37.29	27.87	7.50	35.37	-1.73	-0.08	
8	Post-equated Design 2	2021	27.27	37.33	27.94	7.46	35.40	-1.71	-0.10	-0.02

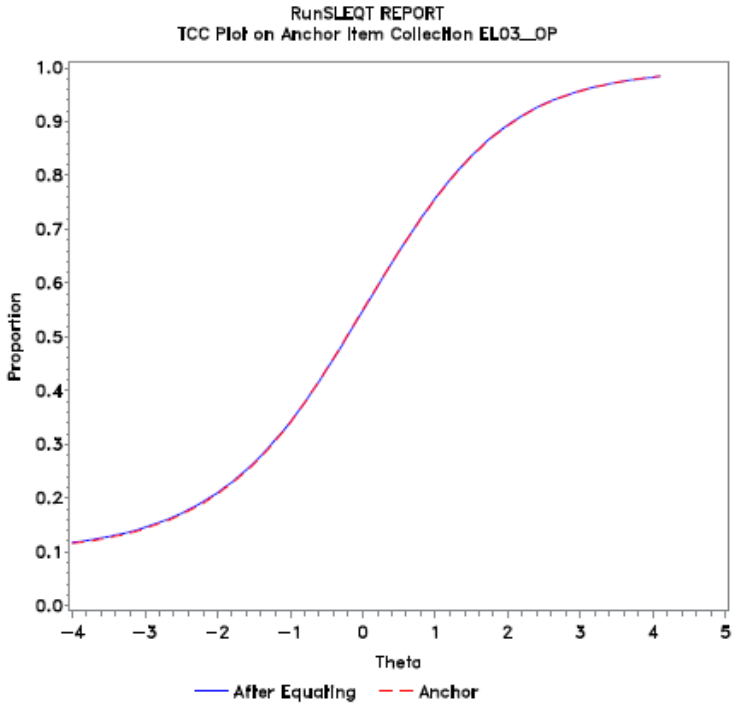
Table 12. Scale Score Summaries, Mathematics

Grade	Design	Year	# of Items	# of Points	N-count	Scale Score Mean	Scale Score Std. Dev.	Median	Min. Score	Max. Score	Mean Difference		
											Spring 2021 - Spring 2019	Spring 2021 Pre-equated - Post-equated	Spring 2021 Design 1 - Design 2
3	Spring 2019	2019	42	42	61151	555.82	53.48	559	360	760			
3	Pre-equated	2021	42	42	30329	545.45	57.91	551	360	760	-10.38		
3	Post-equated Design 1	2021	42	42	30329	545.33	58.04	551	360	760	-10.50	0.12	
3	Post-equated Design 2	2021	42	42	30329	545.89	56.95	551	360	760	-9.93	-0.45	-0.57
4	Spring 2019	2019	46	46	63561	577.14	51.74	582	405	800			
4	Pre-equated	2021	46	46	31400	569.20	54.09	574	405	800	-7.95		
4	Post-equated Design 1	2021	46	46	31400	569.19	53.95	574	405	800	-7.95	0.00	
4	Post-equated Design 2	2021	46	46	31400	569.41	53.62	574	405	800	-7.73	-0.22	-0.22
5	Spring 2019	2019	46	46	64666	601.54	53.08	607	430	830			
5	Pre-equated	2021	46	46	31971	591.41	57.47	598	430	830	-10.13		
5	Post-equated Design 1	2021	46	46	31971	591.39	57.57	598	430	830	-10.15	0.02	
5	Post-equated Design 2	2021	46	46	31971	591.63	57.26	598	430	830	-9.91	-0.22	-0.25
6	Spring 2019	2019	46	46	65393	610.83	58.26	616	440	870			
6	Pre-equated	2021	46	46	32273	599.99	58.56	604	440	870	-10.84		
6	Post-equated Design 1	2021	46	46	32273	599.33	59.34	605	440	870	-11.50	0.66	
6	Post-equated Design 2	2021	46	46	32273	600.09	58.16	605	440	870	-10.75	-0.09	-0.75
7	Spring 2019	2019	46	46	63870	625.39	60.59	632	450	880			
7	Pre-equated	2021	46	46	31604	617.33	61.18	624	450	880	-8.06		
7	Post-equated Design 1	2021	46	46	31604	618.75	59.59	625	450	880	-6.64	-1.42	
7	Post-equated Design 2	2021	46	46	31604	617.82	61.16	625	450	880	-7.57	-0.49	0.93
8	Spring 2019	2019	46	46	62989	644.69	57.68	649	470	890			
8	Pre-equated	2021	46	46	31296	636.89	56.90	640	470	890	-7.80		
8	Post-equated Design 1	2021	46	46	31296	636.13	57.88	640	470	890	-8.56	0.76	
8	Post-equated Design 2	2021	46	46	31296	636.14	57.65	640	470	890	-8.55	0.75	-0.01

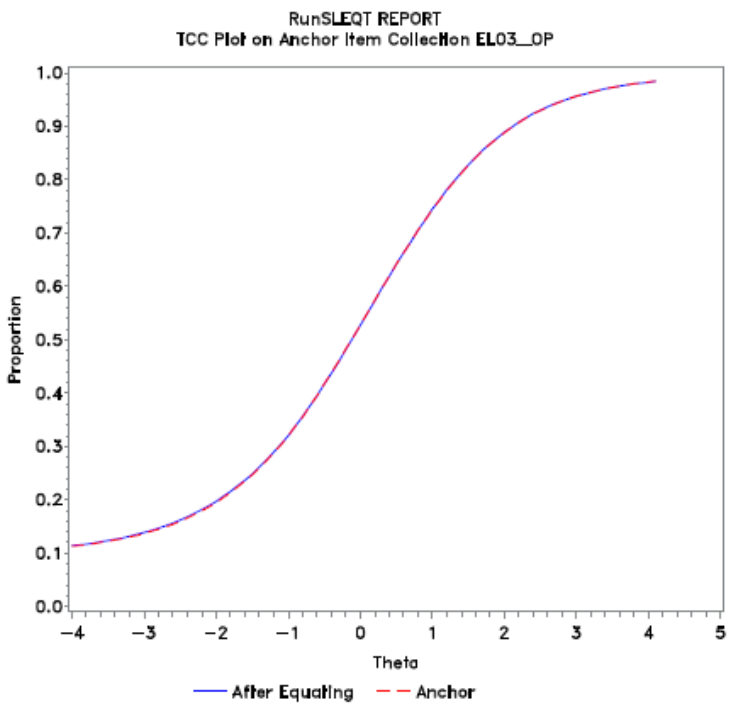
Table 13. Performance Level Summary, Mathematics

Grade	Design	Year	Percentage of Students in Performance Levels					Difference in Percentage of Students Classified as *Proficient* or *Advanced*		
			Below Basic	***Basic***	***Proficient***	***Advanced***	***Proficient* or *Advanced***	**Spring 2021 - Spring 2019**	**Spring 2021 Pre-equated - Post-equated**	**Spring 2021 Design 1 - Design 2**
3	Spring 2019	2019	19.26	31.28	37.18	12.28	49.47			
3	Pre-equated	2021	26.02	31.26	33.02	9.70	42.72	-6.74		
3	Post-equated Design 1	2021	26.07	31.42	32.84	9.67	42.51	-6.96	0.21	
3	Post-equated Design 2	2021	26.02	31.52	32.72	9.74	42.46	-7.00	0.26	0.05
4	Spring 2019	2019	18.82	36.10	32.84	12.23	45.07			
4	Pre-equated	2021	24.10	36.17	30.02	9.71	39.73	-5.35		
4	Post-equated Design 1	2021	24.04	36.17	30.03	9.77	39.80	-5.28	-0.07	
4	Post-equated Design 2	2021	24.17	36.04	30.07	9.73	39.80	-5.28	-0.07	0.00
5	Spring 2019	2019	24.18	29.21	35.11	11.49	46.61			
5	Pre-equated	2021	31.86	28.39	30.54	9.20	39.74	-6.87		
5	Post-equated Design 1	2021	31.64	28.62	30.63	9.11	39.75	-6.86	0.00	
5	Post-equated Design 2	2021	31.60	28.60	30.72	9.08	39.80	-6.81	-0.06	-0.06
6	Spring 2019	2019	26.68	30.80	35.82	6.70	42.52			
6	Pre-equated	2021	34.13	31.74	29.62	4.51	34.13	-8.39		
6	Post-equated Design 1	2021	33.89	31.75	29.99	4.38	34.36	-8.16	-0.23	
6	Post-equated Design 2	2021	33.88	31.77	29.94	4.42	34.36	-8.16	-0.23	0.01
7	Spring 2019	2019	32.10	29.01	34.10	4.79	38.89			
7	Pre-equated	2021	36.91	29.67	30.08	3.34	33.43	-5.46		
7	Post-equated Design 1	2021	36.28	29.74	30.68	3.30	33.99	-4.90	-0.56	
7	Post-equated Design 2	2021	36.59	29.39	30.69	3.33	34.02	-4.87	-0.59	-0.03
8	Spring 2019	2019	28.47	35.63	27.87	8.03	35.90			
8	Pre-equated	2021	33.97	36.73	23.03	6.28	29.30	-6.59		
8	Post-equated Design 1	2021	34.39	36.06	23.44	6.12	29.55	-6.34	-0.25	
8	Post-equated Design 2	2021	34.57	35.84	23.44	6.15	29.59	-6.31	-0.28	-0.04

Figure 1. Anchor Set Input and Estimate TCCs, English Language Arts, Grade 3

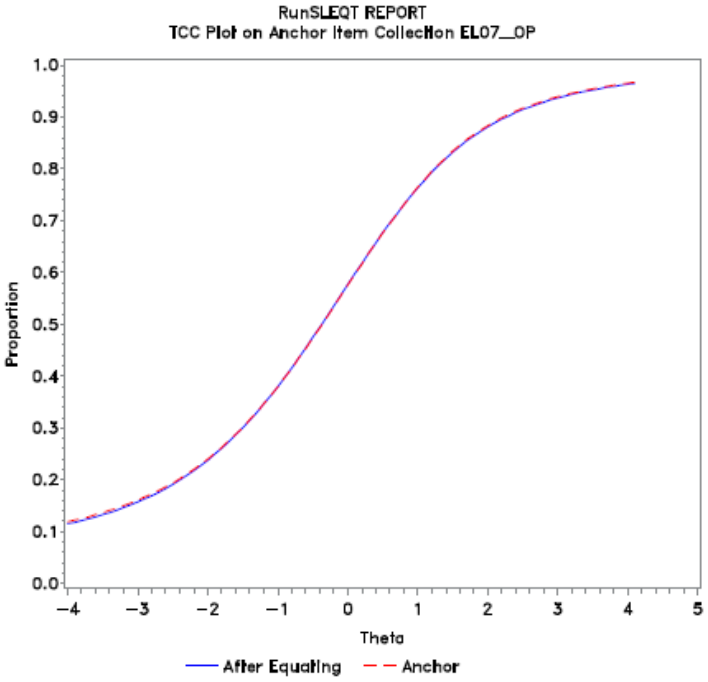


TCCs for Design 1

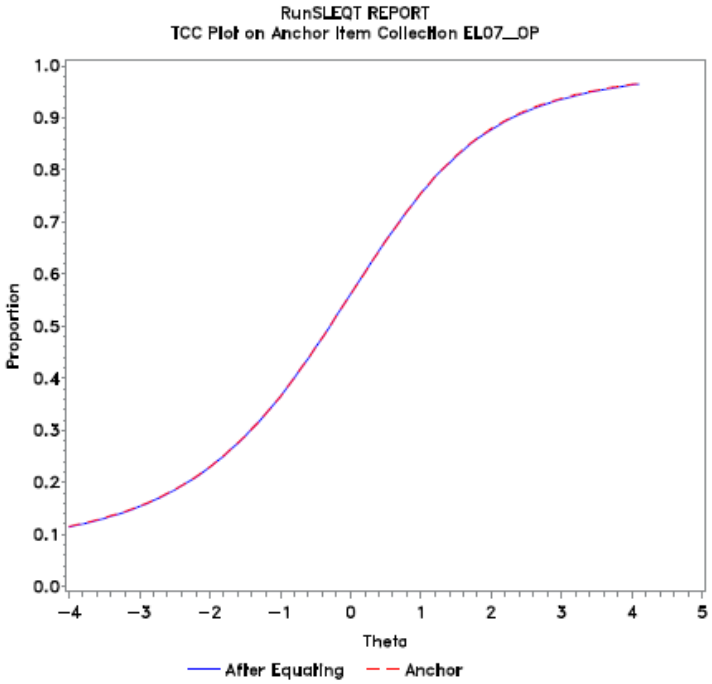


TCCs for Design 2

Figure 2. Anchor Set Input and Estimate TCCs, English Language Arts, Grade 7

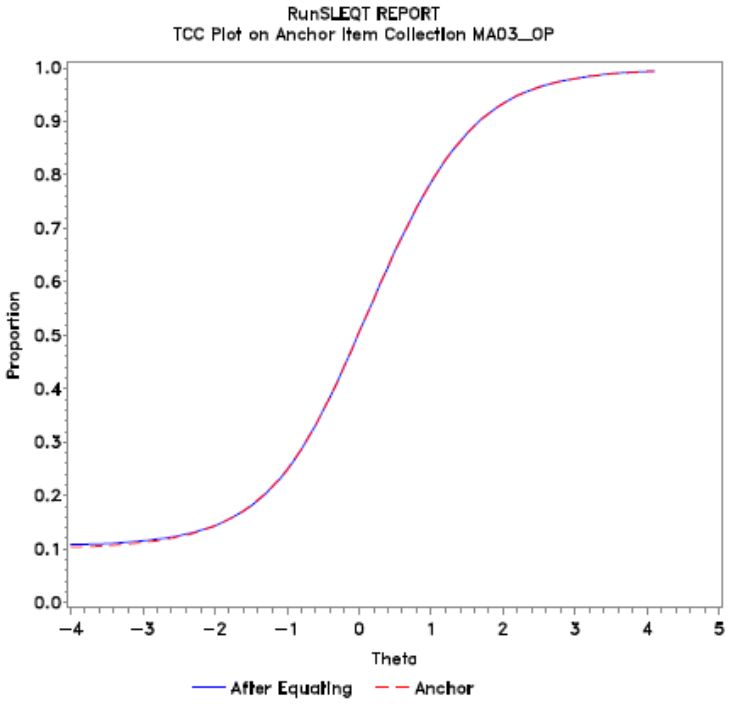


TCCs for Design 1

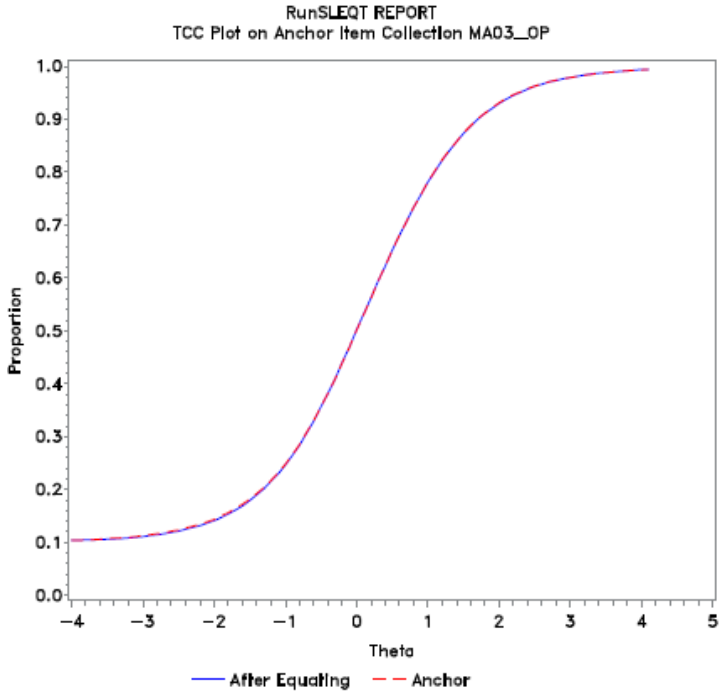


TCCs for Design 2

Figure 3. Anchor Set Input and Estimate TCCs, Mathematics, Grade 3

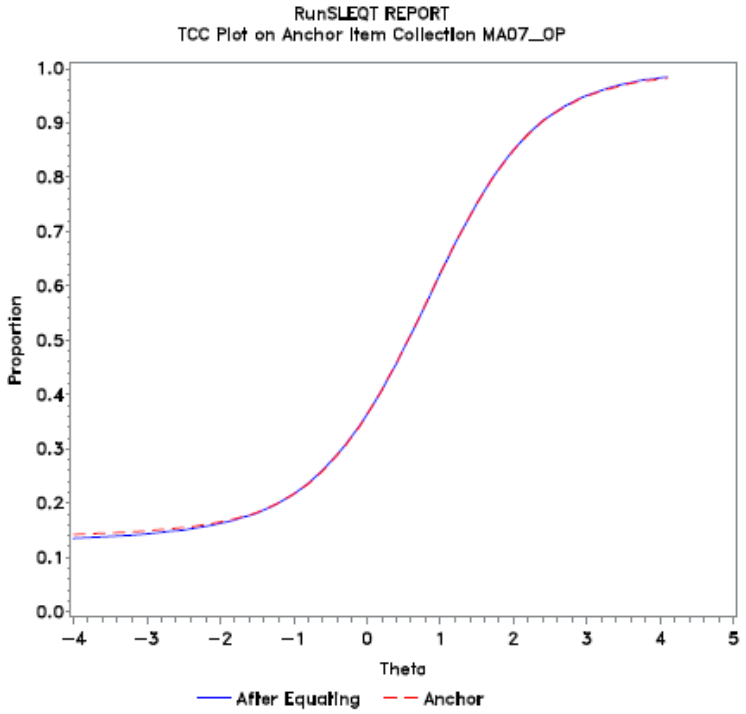


TCCs for Design 1

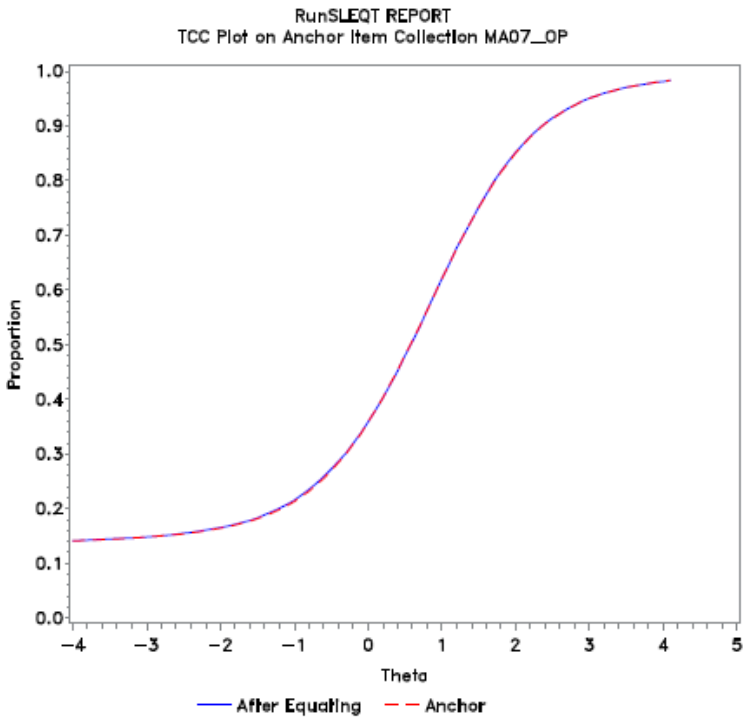


TCCs for Design 2

Figure 4. Anchor Set Input and Estimate TCCs, Mathematics, Grade 7

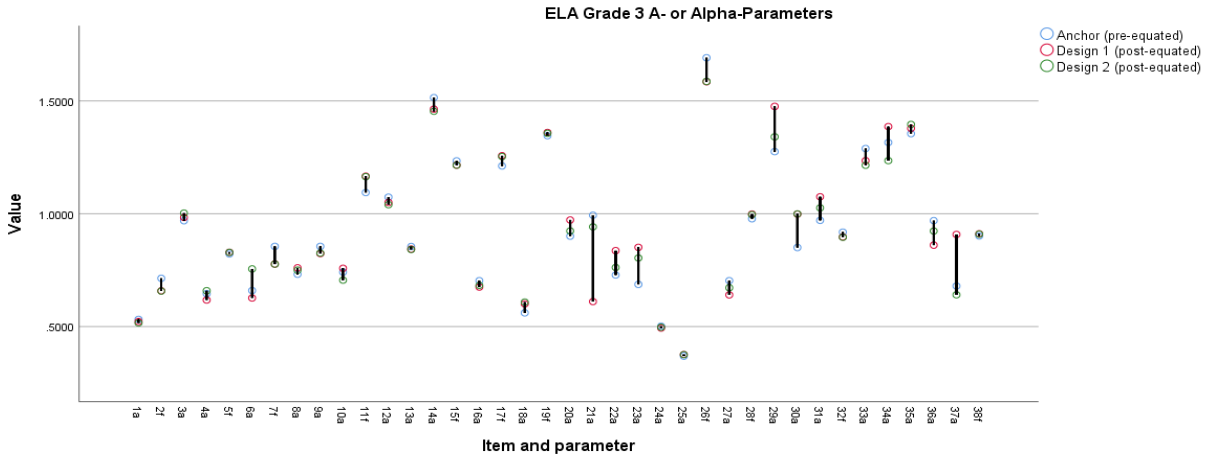


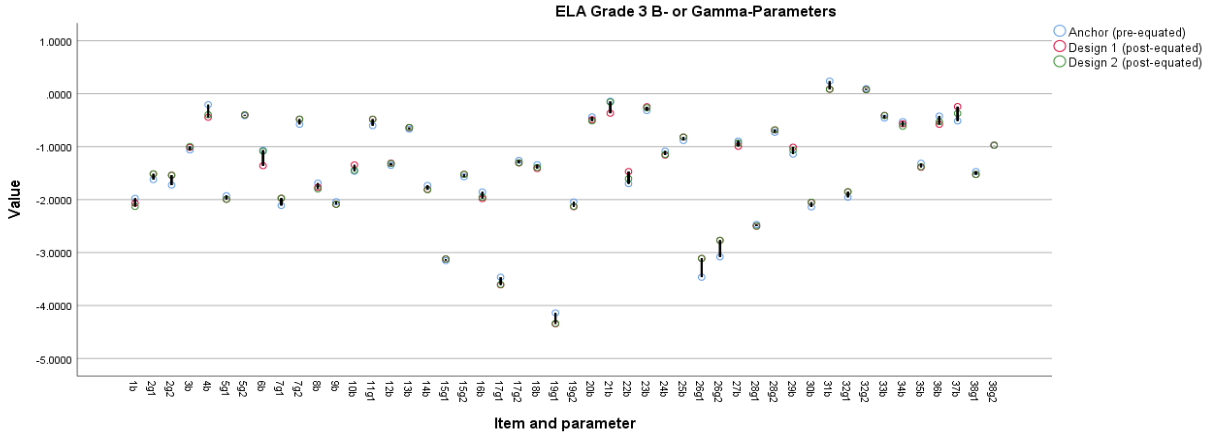
TCCs for Design 1



TCCs for Design 2

Figure 5. Input and Estimate Parameter Values, English Language Arts, Grade 3





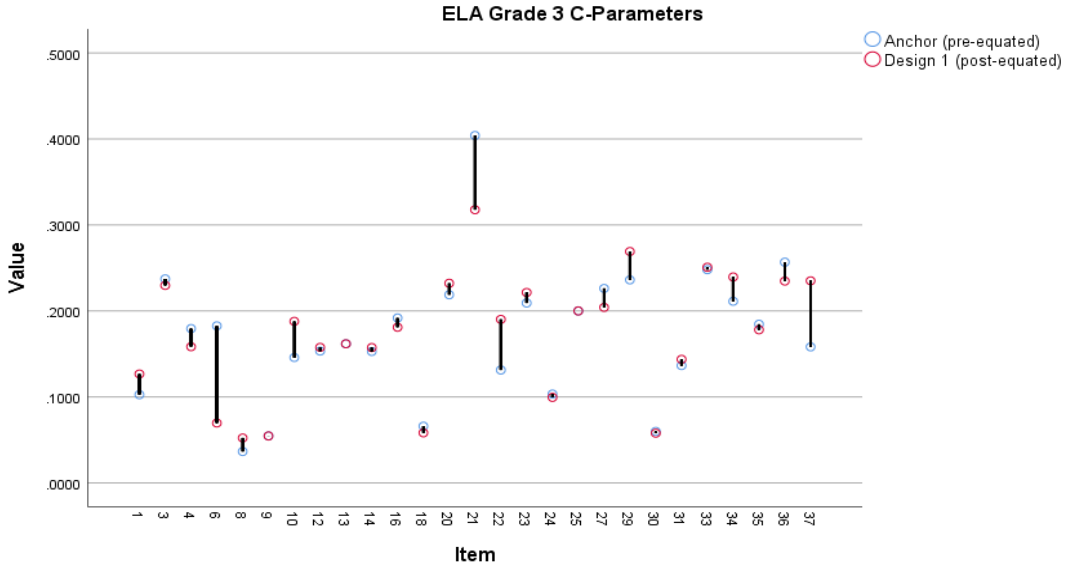
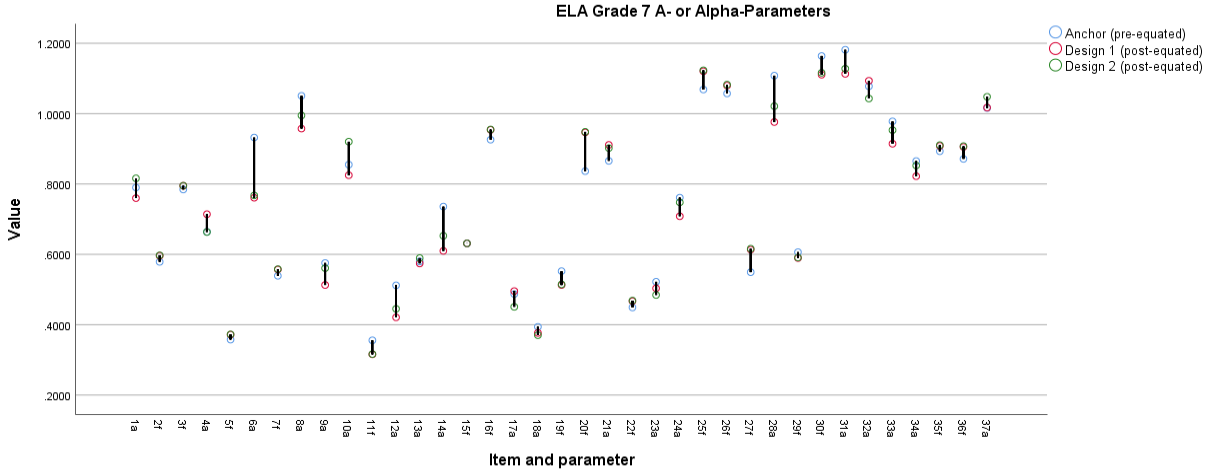
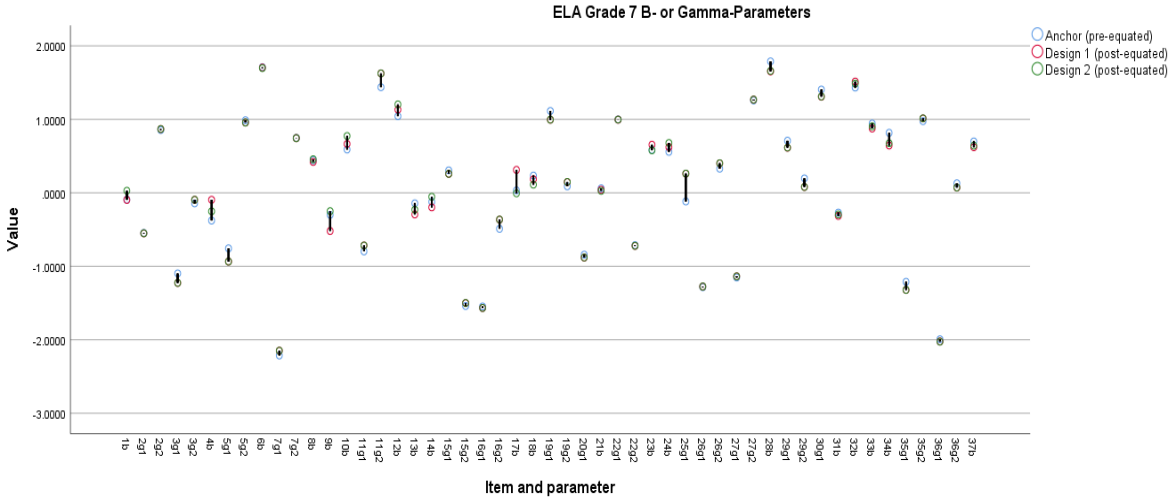


Figure 6. Input and Estimate Parameter Values, English Language Arts, Grade 7





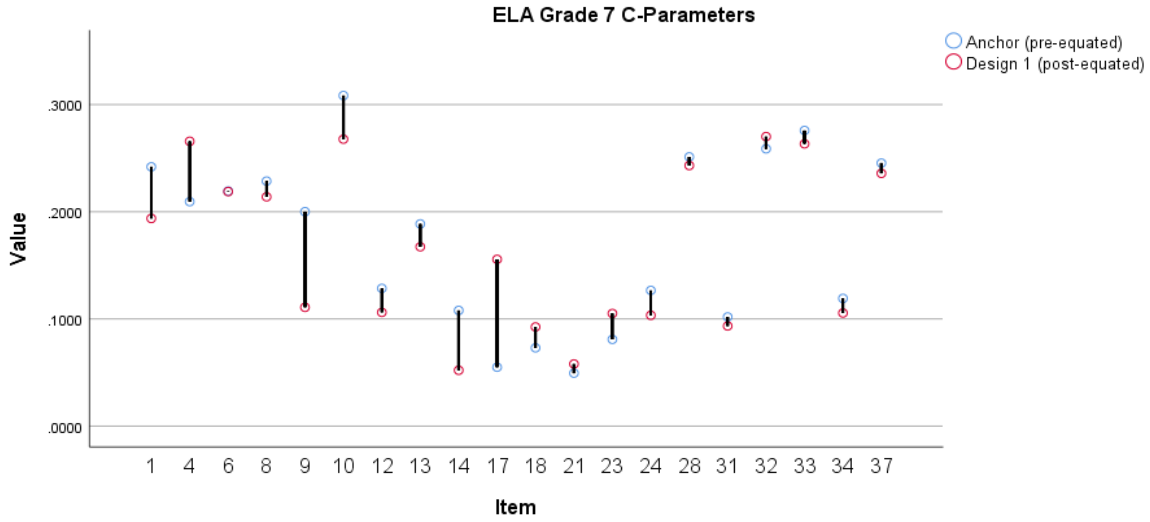


Figure 7. Input and Estimate Parameter Values, Mathematics, Grade 3

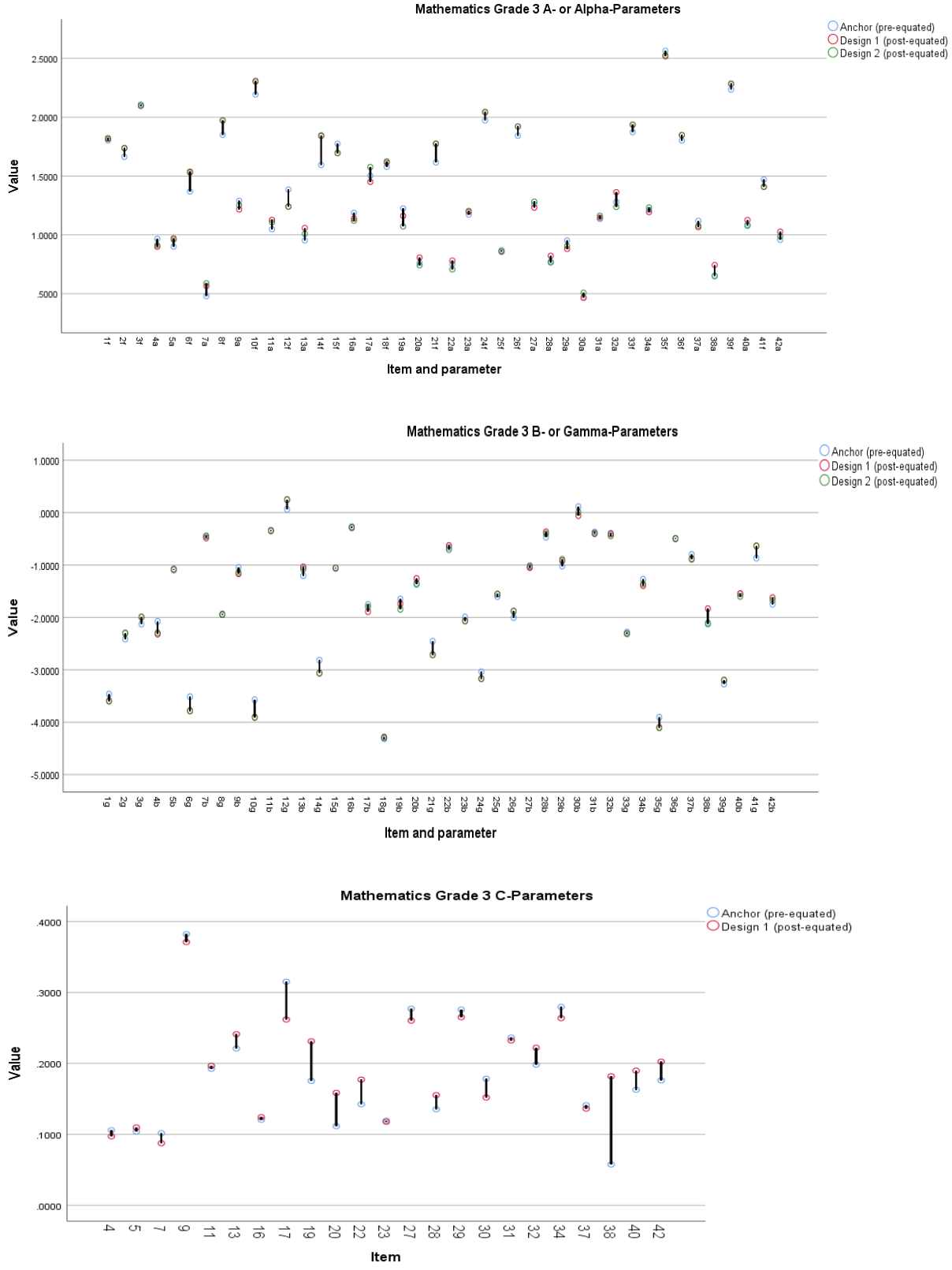


Figure 8. Input and Estimate Parameter Values, Mathematics, Grade 7