2015 No. 009

Findings from the Quality of Items/Tasks/Stimuli Investigations:

PARCC Field Tests

Arthur A. Thacker
Emily R. Dickinson
Bethany H. Bynum
Yao Wen
Erin Smith
Andrea L. Sinclair
Richard C. Deatz
Lauress L. Wise

February 6, 2015

Executive Summary

The Partnership for Assessment of Readiness for College and Careers (PARCC) field tests during the spring of 2014 provided an opportunity to investigate the quality of the items, tasks, and associated stimuli. HumRRO conducted several research studies summarized in this report. Quality of test items is integral to the ***Theory of Action*** (TOA) for PARCC (see Figure 1 for the TOA for PARCC). A TOA indicates the intended uses and expected impact of an assessment system and informs testable claims related to the interpretation of test scores (i.e., the ***Interpretive Argument***). The evidence to support those claims and assumptions represents the ***Validity Argument*** for PARCC. The findings included in this report speak to the validity argument for the Design, Development, Administration, and Scoring phases of PARCC.

Conclusions described here refer to the PARCC field test administration. One purpose of the field test was to gather information about items prior to operational administration. A substantial number of items will be dropped or revised prior to operational testing. This is an expected result of field testing and does not represent a criticism of the operational test. The number and proportion of items that do not meet statistical criteria for some grade subjects, especially for math course assessments, may limit the number of forms that can be constructed with adequate content representation.

Overall, the findings from the research studies on item quality provide the following evidence in relation to the claims from these phases of the PARCC TOA:

Claim 1 (Design): The Assessment Connects with Common Core State Standards (CCSS)

Once we eliminated administratively dropped and poorly functioning items, we examined correlations with state assessment results. Correlations were generally strong and followed expected patterns for convergent/discriminant validity coefficients. This represents one piece of evidence supporting the link between the PARCC assessments and the CCSS. These analyses should be revisited with operational assessments, especially for high school math, where the number of dropped items was large.

Claim 2 (Development): Items are of Sufficient Quality and Rigor

More items than expected were dropped based on administration and statistical quality (item-total correlation)[1]. PARCC may need to develop more items than originally expected during the early years of administration to generate the anticipated number of test forms. No further analyses were conducted on items that were dropped.

[1] Based on typical item survival rates after field test of approximately 85-95% (when items are written by professional item writers).

Item difficulty, which we will use as a proxy for test rigor, was higher than expected. Many items were flagged for very low p-values (few students providing the correct response). This was especially true of math course assessments. Consequently, we recommend focusing item development to ensure that a substantive number of less difficult items are included. We also recommend close monitoring of student classification distributions, overall and by subgroup, and classification accuracy to ensure robust measurement throughout the test scale.

Claim 3 (Administration): Students Respond to Items as Intended

A substantial proportion of students and administrators described issues with the overall test instructions or with item level instructions. Omit rates were high for complex and high-effort item types, especially for high school level students. A substantial proportion of students did not respond as expected to multiple select items, often selecting only one response when asked to select two or more. Some students described confusion related to use of test tools and test navigation. These are things we hope to discover during field test, but they should be surveyed again during operational testing to ensure that these issues have been adequately addressed.

Field test session times were not adequate for several of the sessions. Testing session times were revised based on field-test results, but should be monitored to ensure that students have adequate time to complete each session and to guard against speeded conditions.

Claim 4 (Scoring): Scores Accurately and Reliably Reflect Student Achievement on the Assessed Content

Items requiring handscoring were scored according to PARCC guidelines. Scorers were regularly monitored for consistency and accuracy and retrained or dismissed if they did not reach acceptable levels. We recommend continuing these processes for operational testing and conducting rater drift analyses in the future to ensure consistent scoring within and across years.

Two-part technology enhanced items and evidence-based selected response items with three or more Evidences, as well as two-part technology enhanced items with six or more student responses in the Accuracy portion, were rescored using alternate scoring rules. The alternate scoring rules yielded more normal distributions of student scores, compared to the original rules, which yielded distributions shifted to the left (typically with many 0 scores). The alternate scoring rules did not consistently impact item-total correlations. Investigations of "part-level" scores indicate that item parts may be contributing information on their own. In some instances, one part would function well, but the other poorly, attenuating the information that could be obtained for the item. We recommend that PARCC consider the alternate scoring rules to shift the item-level score distributions toward normal. We also recommend that PARCC consider scoring the item parts separately to optimize test information and to allow for the omission of a poorly performing part, without losing the entire item.

Cautions

The studies described in this report represent HumRRO's investigations. ETS is also conducting numerous psychometric evaluations of the PARCC field test items. HumRRO and ETS strived to avoid duplication of effort. This report and ETS' report related to the psychometric properties of PARCC field test items should be considered in tandem.

Table of Contents

List of Tables

List of Figures

Findings from the Quality of Test Items/Tasks/Stimuli Investigations: PARCC Field Tests

Purpose

The Partnership for Assessment of Readiness for College and Careers (PARCC) conducted field tests in Math and English Language Arts (ELA)/Literacy during the spring of 2014. In approximately March-April 2014 Performance Based Assessments (PBA) were administered; in May-June 2014 End-of-Year (EOY) assessments were conducted. These field tests provided opportunities for initial investigations of the quality of the items, tasks, and associated stimuli. The quality of test items is integral to the ***Theory of Action*** (TOA) for PARCC. A TOA indicates the intended uses and expected impact of an assessment system and informs testable claims related to the interpretation of test scores (i.e., the ***Interpretive Argument***). The evidence to support those claims and assumptions represents the ***Validity Argument*** for PARCC. The findings included in this report speak to the validity argument for the Design, Development, Implementation, Administration, and Scoring phases of PARCC. The *Standards for Educational and Psychological Testing* (hereafter referred to as the *Standards*) were jointly authored by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education and provide criteria for the development and evaluation of tests and testing practices as well as guidelines for assessing the validity of interpretations of test scores for the intended test uses (AERA, APA, & NCME, 2014, p. 1). Throughout this report references are made to findings that relate to the administration criteria from the *Standards*.

Background

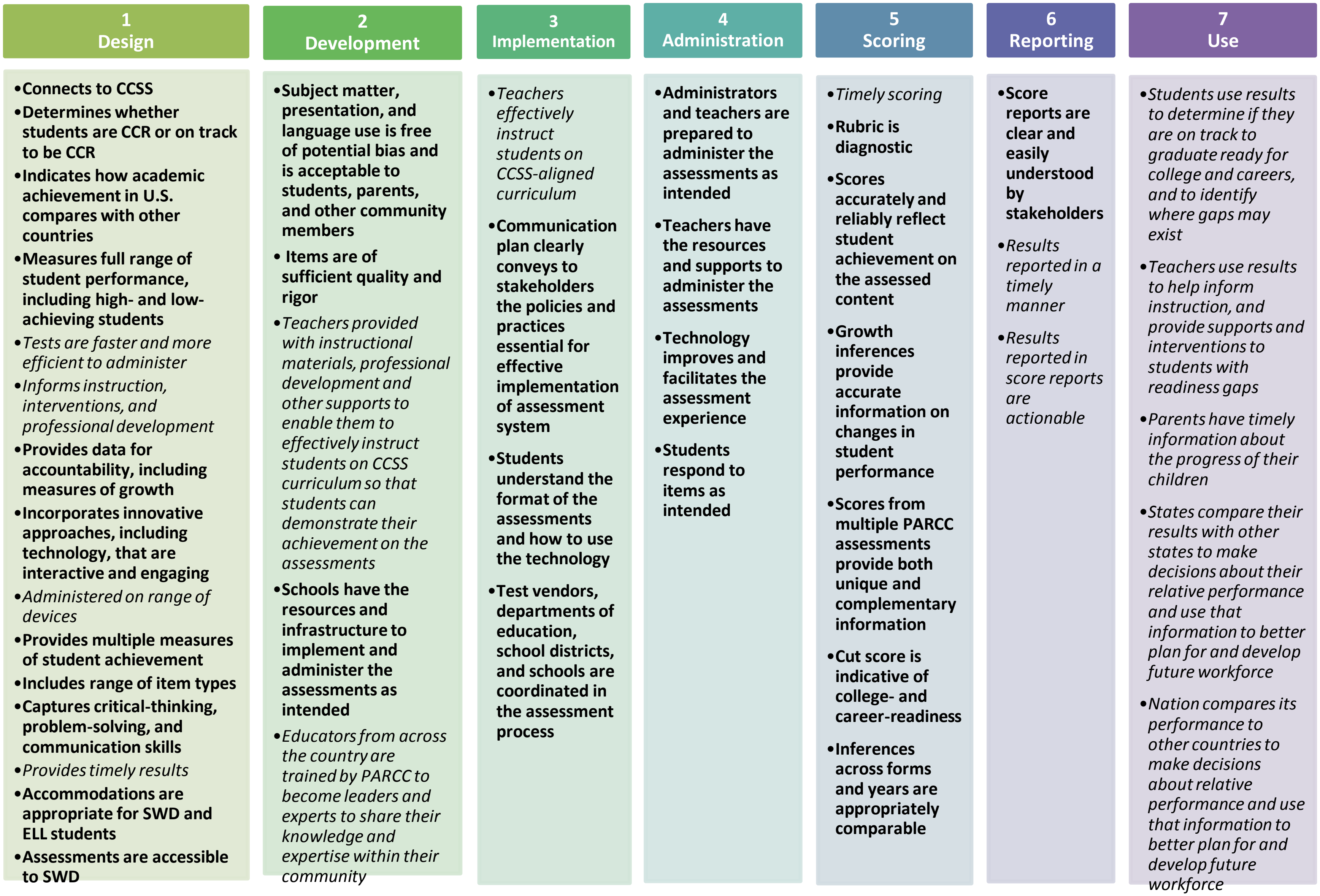
HumRRO's studies of the PARCC field tests take an argument-based approach to validation, whereby the goal is to evaluate the extent to which the evidence supports or refutes the proposed interpretations and uses of test scores (Kane, 2006). This approach is consistent with the *Standards* on validity, which state that, "a rationale should be presented for each intended interpretation of test scores for a given use," and that that rationale "should indicate what propositions are necessary to investigate the intended interpretation" of test scores (AERA, APA, & NCME, 2014, p. 23). The *Standards* caution against using the unqualified phrase "the validity of the test." Rather, validity is a matter of the validity of the interpretations and uses of test *scores*. Consequently, even though scores are not reported on the PARCC field tests, the intended interpretation and use of scores is what drives early validation efforts. Therefore, it is appropriate and necessary to refer to threats to the validity of the intended interpretations and uses of PARCC scores, even though PARCC scores are not reported for the field tests. That scores are not reported for the field tests does not change the intended interpretation and use of PARCC scores.

PARCC scores are intended to be interpreted and used as an indication of whether students are on track in their learning to be successful in college and their careers[2]. This is commonly referred to as “college and career readiness,” or CCR. Based on the stated claims and purposes of the PARCC assessments as described in PARCC’s publicly available documentation, a TOA for PARCC was developed with the end goal of college and career readiness (see PARCC Validity Studies Memorandum; Thacker, Sinclair, Wise, & Becker, 2014). The TOA for PARCC is organized as a series of interim goals that lead to this end goal, as depicted in Figure 1. Again, the end goal of the PARCC assessment system is that PARCC scores in ELA and mathematics provide an indication that students are, or are not, college and career ready (CCR), or on track to become CCR for younger students. In order to attain the end goal of college and career readiness, each of the interim goals (denoted by the column headers in Figure 1) must be met. Lack of support for any of the interim goals undermines the validity of the system to meet its end goal.

This report focuses on the PARCC items, tasks, and associated stimuli, heretofore referred to as *items* for simplicity. In a sense, items are the raw materials of tests. Therefore, the characteristics of items have implications throughout the assessment cycle. The findings in this report address several interim goals in the PARCC TOA. Each goal is met by establishing support for *claims* that have been designated as foundational to that goal. Those claims are depicted by the bullet points under the interim goal headings in the TOA. The interim goals and associated claims addressed by this study are: a) Design: the assessment connects with Common Core State Standards (CCSS), b) Development: items are of sufficient quality and rigor, c) Administration: students respond to items as intended, and d) Scoring: scores accurately and reliably reflect student achievement on the assessed content.

Numerous assumptions underlie each of these claims. For example, if the PARCC assessment connects with the CCSS, then scores on PARCC would be expected to correlate positively with other assessments that are designed to measure CCSS knowledge and skills (AERA, APA, NCME, 2014, pp. 16-17). The claims identified above served as a guide to HumRRO for identifying assumptions that must be met for those claims to be true. Those assumptions relate to the research questions guiding the investigations of the PARCC item quality. The research questions and the investigations designed to address those questions were proposed by HumRRO to PARCC and revised after multiple rounds of feedback from PARCC leadership.

[2] From PARCC website: http://www.parcconline.org/about-parcc.



1 Design

- **Connects to CCSS**
- **Determines whether students are CCR or on track to be CCR**
- **Indicates how academic achievement in U.S. compares with other countries**
- **Measures full range of student performance, including high- and low-achieving students**
- *Tests are faster and more efficient to administer*
- *Informs instruction, interventions, and professional development*
- **Provides data for accountability, including measures of growth**
- **Incorporates innovative approaches, including technology, that are interactive and engaging**
- *Administered on range of devices*
- **Provides multiple measures of student achievement**
- **Includes range of item types**
- **Captures critical-thinking, problem-solving, and communication skills**
- *Provides timely results*
- **Accommodations are appropriate for SWD and ELL students**
- **Assessments are accessible to SWD**

2 Development

- **Subject matter, presentation, and language use is free of potential bias and is acceptable to students, parents, and other community members**
- **Items are of sufficient quality and rigor**
- *Teachers provided with instructional materials, professional development and other supports to enable them to effectively instruct students on CCSS curriculum so that students can demonstrate their achievement on the assessments*
- **Schools have the resources and infrastructure to implement and administer the assessments as intended**
- *Educators from across the country are trained by PARCC to become leaders and experts to share their knowledge and expertise within their community*

3 Implementation

- *Teachers effectively instruct students on CCSS-aligned curriculum*
- **Communication plan clearly conveys to stakeholders the policies and practices essential for effective implementation of assessment system**
- **Students understand the format of the assessments and how to use the technology**
- **Test vendors, departments of education, school districts, and schools are coordinated in the assessment process**

4 Administration

- **Administrators and teachers are prepared to administer the assessments as intended**
- **Teachers have the resources and supports to administer the assessments**
- **Technology improves and facilitates the assessment experience**
- **Students respond to items as intended**

5 Scoring

- *Timely scoring*
- **Rubric is diagnostic**
- **Scores accurately and reliably reflect student achievement on the assessed content**
- **Growth inferences provide accurate information on changes in student performance**
- **Scores from multiple PARCC assessments provide both unique and complementary information**
- **Cut score is indicative of college- and career-readiness**
- **Inferences across forms and years are appropriately comparable**

6 Reporting

- **Score reports are clear and easily understood by stakeholders**
- *Results reported in a timely manner*
- *Results reported in score reports are actionable*

7 Use

- *Students use results to determine if they are on track to graduate ready for college and careers, and to identify where gaps may exist*
- *Teachers use results to help inform instruction, and provide supports and interventions to students with readiness gaps*
- *Parents have timely information about the progress of their children*
- *States compare their results with other states to make decisions about their relative performance and use that information to better plan for and develop future workforce*
- *Nation compares its performance to other countries to make decisions about relative performance and use that information to better plan for and develop future workforce*

Figure 1. Seven sequential stages from the simplified TOA for the PARCC assessment system.

Note. **Bold** font indicates claims related to test score interpretation and *italics* indicate claims regarding impact/consequences of test scores.

Those research questions and the relevant *Standards* that they address are:

1. Do items and tasks measure the intended CCSS knowledge and skills (Standard 3.2)?
2. Do students interact with the items/tasks and stimuli as intended given the instructions (Standard 4.16)?
3. What is the optimal testing time required for individual or sets of items/tasks to allow students to demonstrate what they know and can do with regard to the target of measurement (Standard 4.14)?
4. Do rubrics lend themselves to accurate and reliable scoring (Standards 4.20, 4.21, and 4.23)?

The above list of research questions and the assumptions they address should not be considered an exhaustive list of research questions on item quality. Rather, these research questions, and the claims they address, represent a combination of the topics of greatest interest at the time of PARCC field test administration and those that are most feasible on which to collect information. As stated in the PARCC validity studies memorandum (Thacker et al., 2014), the validity argument should be updated annually to document the *continuing* collection of validity evidence for the PARCC assessments. Furthermore, if support for a claim is weak or the evidence is inconclusive, then that aspect of the testing program must be improved or additional evidence gathered. As such, the information reported here should not be treated as the final word on the validity argument regarding PARCC item quality.

Relevant Standards from the *Joint Standards*

Standard 3.2: Test developers are responsible for developing tests that measure the intended construct…

Standard 4.14: For a test that has a time limit, test development research should examine the degree to which scores include a speed component and should evaluate the appropriateness of that component, given the domain the test is designed to measure.

Standard 4.16: The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended…

Standard 4.20: The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale…

Standard 4.21: When test scores are responsible for scoring and scoring requires scorer judgment, the test user is responsible for providing adequate training and instruction to the scorers and for examining scorer agreement and accuracy. The test developer should document the expected level of scorer agreement and accuracy and should provide as much technical guidance as possible to aid test users in satisfying the standard.

Standard 4.23: When a test score is derived from the differential weighting of items or subscores, the test developer should document the rationale and process used to develop, review, and assign item weights.

The next section of the report outlines the method used to collect evidence for the research questions on PARCC item quality.

Data Sources

This section is organized around the four data sources used to address the four research questions. These include PARCC test maps and student data files provided by Pearson, state test scores from a sample of PARCC states, student and Test Administrator survey data, and site visit data collected by HumRRO.

PARCC Test Maps and Student Data Files

Pearson, the PARCC field test assessment contractor, provided two sets of test map files to HumRRO. The first set of test maps, referred to as the *short test maps*, included a record for each item. Each record included response type, interaction type, Common Core State Standard, and PARCC sub-claim, among other item-level information. The second set of test maps, referred to as the *long test maps*, included a record for each item part. In other words, if an item included two parts, the long test map included a record for the overall item and each sub-part (three records total). These files also included information on the item type, segment, XML scoring information, and maximum score points. Separate test maps were provided for each grade, subject, and test form. We combined all long test maps together across grade and subjects, then did the same for all short test maps. We then merged the two files together using Item ID (UIN, Entity ID). Separate master files were constructed for PBA and EOY. Grade- and subject-specific files were created using the grade (grade level, FORMSOUTPUT_3, Starting Grade) and subject (subject, subtest). Answer keys were constructed by pulling number of responses prompted for and associated correct responses from the xml scoring information. We ensured our test maps included all items administered on the test by comparing the item IDs in the student data files to the constructed test maps.

Several student data files were analyzed. First, Pearson provided PBA and EOY student data files for each grade, subject and session type (C1 and C2). All student data files were provided via password protected secure file transfer protocol (SFTP) site. For each student, these files included demographic information, the list of items taken, total score, subscores, item-level scores, item response data, rater scores, item-part responses (for choice interaction single and multiple cardinality items only), and information on whether students omitted an item (SIRB). The student data files are organized such that item level scores correspond to the book position of the item for a particular form. For example, if two students received two different forms, then Item 1 listed in the student data file may correspond to two different items. In essence, each record in this file represented one student's experience, including all the items in the order they were presented. Our analyses necessitated reconfiguring the file so that all responses to a specific test item were recorded in the same variable for all students. For a number of our analyses, it was necessary to reconstruct the data so that scores on the same item (regardless of the form taken) were listed as a unique variable. We used the test maps to construct a student data matrix where every item administered for each grade and subject was a separate variable and student scores for those that took the item is reported in the variable and missing for those that did not take the item.

Pearson also provided a student response data file for all tested students (across grades and subjects). This data file contained student and school IDs, form name, item IDs, and complete student responses to all items, including subparts of items. Pearson coded some items as "Do Not Use (DNU)." DNU items were excluded for statistical performance or other reasons (content review, etc.). Unless items were coded DNU, they were included in subsequent analyses[3]. This data file was merged with the student data file and used for two analyses using student responses to part items.

State Assessment Data

HumRRO received recommendations from PARCC staff regarding states to solicit for 2014 state assessment scores. States were recommended, in part, based on the number of students who participated in the PARCC assessment, with a preference for states with larger numbers of participating students. Of the 16 PARCC states, six were sent data requests, and five of those states provided data.

Several steps were taken to facilitate the secure delivery of state assessment scores. First, HumRRO accessed PARCC registration information via PearsonAccess and assigned each registered student across participating states a randomly generated, unique identification number. Files containing the state assigned student ID and the HumRRO-generated ID were then sent to those states agreeing to provide data, allowing states to merge in student scores without requiring other identifying information. The states uploaded final files containing the HumRRO-generated ID with associated state ELA and Math scores to HumRRO's secure ftp site.

State assessment scores were merged into student data files provided by Pearson using the state student ID as the matching variable. In order to maintain the anonymity of participating states, the numbers of student records from each state are not reported. However, the numbers of students used in particular analyses are reported in subsequent sections of the report where appropriate. Also, because state assessments vary in their reporting scales, all comparisons between PARCC and state assessments were conducted separately for each state.

[3] This may include a small number of "off-grade" items placed on a small number of forms for a linking study. If they were part of a form and not coded as DNU, they were included.

Student and Test Administrator Survey Data

HumRRO developed surveys to capture information about students' and test administrators' experiences with the PARCC assessments. This section provides brief descriptions of the procedures associated with each survey. Complete information regarding the survey instruments and respondents is available in HumRRO's report on the quality of PARCC test administration (Sinclair, Deatz, Johnston-Fisher, Levinson, & Thacker, 2014).

Student surveys. Upon completion of the PBA and EOY field tests, students completed a survey about their test-taking experience. Different surveys were administered following completion of math and ELA assessments. There was considerable overlap in the content of the items on the math and ELA surveys, although each contained a subset of items specific to the content area being assessed. There was no difference in the student surveys for PBA and EOY, although the survey data was collected and reported separately for each. Slightly edited versions of the surveys were administered to students who were administered paper-based versions of the field tests.

Test Administrator surveys. An anonymous online Test Administrator (TA) Survey was completed by TAs following administration of the PARCC PBA and/or EOY field tests. The primary purpose of the survey was to collect information from the perspective of TAs on the effectiveness of the training they received and to identify potential problems that arose during test administration. The TAs were provided a web link to the online survey in their test administrator manual. They were instructed to log on and complete the survey after test administration was completed. If TAs administered both the PBA and the EOY, then they were instructed to complete the survey *after* the EOY administration, only. The TAs responded to a separate survey for computer/tablet-based test administration and for paper-based test administration. The online test administrator survey was available from March 24, 2014 to June 13, 2014, which corresponded to the first day of PARCC field test administration and to one week after the last day of PARCC field test administration.

Scoring Center Site Visits

HumRRO staff conducted observations of scoring centers to evaluate the constructed response hand-scoring process and to develop an observation data collection tool for use during operational scoring. HumRRO adapted established observation protocols used for the National Assessment for Educational Progress (NAEP) for over 10 years. The protocols were revised and refined during the summer 2013 PARCC item tryout scoring.

The protocols consist of two checklists with expected scoring tasks or activities related to either scorer training or item scoring. Trained observers check all of the tasks they observe during their visit. The original NAEP checklists were revised for PARCC based on the training and scoring tasks in the PARCC Handscoring Specification 4.1 document. The specifications document provides process information for scorer training and qualification, student response backreading and scorer calibrations, staff qualifications, and project clean-up activities (item closeout). Additionally the specifications include quality measure requirements for validity checkpoints, rangefinding, and inter-rater reliability (IRR) standards.

HumRRO visited Pearson Educational Measurement in Virginia Beach three times; the first was in July 2013 for the PARCC item tryout scoring and the other two were in June 2014 for field test scoring. The purpose of the item tryout is to document how test items or tasks function, and to refine processes and procedures accordingly. For scoring in particular, item tryouts provided an opportunity to pilot scoring processes, including the use of the ePen image-based scoring system, and to identify score exemplars and refine rubrics as needed. For HumRRO's purposes, observing the item tryout scoring provided an opportunity to try out the observation checklists and refine them before subsequent visits.

The item tryout site visit (17–18 July 2013) began with a WebEx teleconference led by key Pearson project leaders in Iowa and conducted for the Virginia Content Specialists and Scoring Directors. The purpose of the meeting was to provide an update on the PARCC scoring plans and to describe the item tryout process beginning that week. Following the meeting, Scoring Directors (in the role of scorers) paired with Content Specialists for the item tryout scoring. Scorer training was abbreviated because seasoned Scoring Directors were used; however, scoring processes were not.

After the item tryout scoring observation, the observation checklists for scorer training and item scoring were revised. These Excel® files allow observers to record qualitative observation data and provide one-click access to resources such as information on the minimum requirements for quality statistics, backreading, and validity checks, as well as the scoring rubric. The observer is prompted to note when system parameters are exceeded (e.g., inter-rater agreement statistics, scoring and productivity rates) and if the Scoring Director initiates any individual, or group, scorer training interventions as a result. The modified checklists were used during the two site visits in June 2014, for scorer training and item scoring (see Appendix A).

Analyses and Results

The following sections describe the analyses and results for each of our four research questions. These analyses do not represent the complete item quality analyses conducted based on the PARCC field test. Educational Testing Services (ETS) served as the primary psychometric processors for the field test and conducted several item quality analyses in that capacity. HumRRO and ETS worked in cooperation and made efforts not to unnecessarily replicate analyses. ETS' reports should be considered in concert with this report to generate a full understanding of all of the item quality analyses conducted for the PARCC field test.

Findings for Research Question 1: Do items and tasks measure the intended CCSS knowledge and skills?

Two sets of analyses were conducted to address research question 1. First, item analysis statistics were calculated to identify any potentially problematic items. Next, correlations with state assessment scores for a subsample of participating states were run to assess convergent and discriminant validity evidence.

Summary of item statistics. A key purpose of field testing any assessment is to evaluate the quality of the item pool that has been generated to measure the construct or trait of interest[4]. Items with statistical properties that fall outside of acceptable ranges will cause problems during IRT model estimation, and so such items will either need to be edited and subjected to additional field testing, or will need to be removed from the item pool. Losing a large number of items would raise concerns about the adequacy of the item pool for creating parallel test forms that measure the intended CCSS knowledge and skills. This would be especially problematic if these poorly functioning items were concentrated in particular sub-topics within a grade/subject.

The initial data files contained codes indicating that particular items had already been identified by Pearson as items that should be dropped due to presentation or scoring issues (coded DNU, as described previously). The number and percentage of these 'administrative drops' are summarized in Table 1. Among the EOY assessments, between 10% (Grade 3 math) and 28% (Integrated Math 2) of items were identified for removal prior to the evaluation of item statistics. Among the PBA assessments, between 12% (Grade 9 ELA) and 42% (Integrated Math 3) of items were identified for removal prior to the evaluation of item statistics.

Student data on the remaining items were analyzed to determine item p-values and item-total correlations. For dichotomously scores items, p-values reflect the proportion of students answering the item correctly. For polytomously scored items, p-values reflect the proportion of the maximum possible score points on the items. The p-values indicate how easy or hard an item is; extremely easy or hard items have limited value in the assessment.

[4] This section reflects the rescore of the field test data completed by Pearson in early 2015. Relatively few items (fewer than 200 across all grade/subjects) were impacted by the rescore.

Table 1. Administered Item Information and Dropped Items by Test, Subject, and Grade

Test	Subject	Grade	Total Items Administered	Items After Administrative Drops	# Drops Administrative	% Drops Administrative
EOY	ELA	3	307	239	68	22%
	ELA	4	319	280	39	12%
	ELA	5	293	246	47	16%
	ELA	6	316	261	55	17%
	ELA	7	303	242	61	20%
	ELA	8	294	233	61	21%
	ELA	9	315	252	63	20%
	ELA	10	323	266	57	18%
	ELA	11	311	278	33	11%
	Algebra 1	--	501	401[†]	100	20%
	Algebra 2	--	436	344	92	21%
	Geometry	--	485	400	85	18%
	Int. Math 1	--	213	171	42	20%
	Int. Math 2	--	173	125[†]	48	28%
	Int. Math 3	--	197	143	54	27%
	Math	3	534	479	55	10%
	Math	4	540	458	82	15%
	Math	5	518	437	81	16%
	Math	6	507	385	122	24%
	Math	7	455	395	60	13%
	Math	8	504	431	73	14%
PBA	ELA	3	243	186	57	23%
	ELA	4	281	240	41	15%
	ELA	5	250	200	50	20%
	ELA	6	281	228	53	19%
	ELA	7	229	199	30	13%
	ELA	8	243	202	41	17%
	ELA	9	314	277	37	12%
	ELA	10	260	212	48	18%
	ELA	11	280	243	37	13%
	Algebra 1	--	307	209	98	32%
	Algebra 2	--	314	198	116	37%
	Geometry	--	312	224	88	28%
	Int. Math 1	--	82	62	20	24%
	Int. Math 2	--	85	57	28	33%
	Int. Math 3	--	91	53	38	42%
	Math	3	292	236	56	19%
	Math	4	311	245	66	21%
	Math	5	284	213	71	25%
	Math	6	284	206	78	27%
	Math	7	288	233	55	19%
	Math	8	322	252	70	22%

Note. ELA = English Language Arts.

[†] One item had no student responses after drops were performed and was excluded from further analyses.

Item-total correlations are correlations between the score on individual items with that of the total test. Identifying an item that performs differently from the test as a whole is an indication of a potentially problematic item: possibly poorly written, scored incorrectly, or a measure of a construct that differs substantially from the rest of the test. HumRRO flagged items using criteria provided by ETS. All item-total correlations are based on a single form (students' scores on an individual item correlated with their scores on the full test form on which the item appeared, PBA and EOY treated as distinct forms). Dichotomous items with p-values greater than .95 and polytomous items with p-values greater than .80 were flagged. P-values below .30 were also flagged, as were items with item-total correlations that were less than .20. HumRRO dropped items with item-total correlations less than 0.20 from further analyses. Table 2 summarizes the percentage of items that flagged on each criterion. The number (or percentage) of administrative drops from Table 1 can be added to the number (or percentage) of statistical drops in Table 2 to create a total number of items dropped from further analyses[5].

Table 2. Item Statistics across Forms after Administrative Drops by Test, Subject, and Grade

			Item Difficulty (p-value)		Item Total Correlation		Statistical Drops Identified by HumRRO		
Test	**Subject**	**Grade**	**p-value > .95 or > .80**	**p-value < .30**	**ITC < .20**	**Negative ITC**	**# Stat Drops**	**% Stat Drops**	**Final # of Items used in Analysis**
EOY	ELA	3	0%	38%	3%	0%	8	3%	231
	ELA	4	0%	31%	2%	0%	6	2%	274
	ELA	5	0%	29%	4%	0%	11	4%	235
	ELA	6	0%	28%[‡]	3%	0%	9	3%	252
	ELA	7	1%	21%	5%	0%	11	4%	231
	ELA	8	1%	25%	6%	0%	14	5%	219
	ELA	9	0%	37%	6%	0%	16	5%	236
	ELA	10	0%	37%	4%	0%	9	3%	257
	ELA	11	0%	47%	2%	0%	6	2%	272
	Algebra 1	--	0%	68%	31%	1%	75	15%	326
	Algebra 2	--	0%	64%	22%	0%	42	10%	302
	Geometry	--	0%	67%	9%	0%	35	7%	365
	Int. Math 1	--	0%	63%	17%	0%	29	14%	142
	Int. Math 2	--	0%	74%[‡]	32%	4%	25	14%	100
	Int. Math 3	--	0%	71%[‡]	24%	1%	20	10%	123
	Math	3	1%	21%	5%	0%	24	4%	455
	Math	4	0%	18%	3%	0%	13	2%	445
	Math	5	0%	29%	3%	1%	13	3%	424
	Math	6	0%	39%	5%	0%	20	4%	365
	Math	7	0%	55%	10%	0%	41	9%	354
	Math	8	0%	62%	6%	0%	28	6%	403
	ELA	5	**	**	6%	0%	11	4%	189

(continued)

[5] The percentage of items flagged (p < .20) does not always match the statistical drops because items may have been flagged on multiple forms and counted each time, but only counted once as a dropped item.

Table 2. Item Statistics across Forms after Administrative Drops by Test, Subject, and Grade (continued)

Test	Subject	Grade	Item Difficulty (p-value)		Item Total Correlation		Statistical Drops Identified by HumRRO		
			p-value > .95 or > .80	p-value < .30	ITC < .20	Negative ITC	# Stat Drops	% Stat Drops	Final # of Items used in Analysis
PBA	ELA	3	**	**	3%	0%	6	2%	180
	ELA	4	**	**	5%	0%	12	4%	228
	ELA	6	**	**	2%	0%	4	1%	224
	ELA	7	**	**[‡]	3%	1%	5	2%	194
	ELA	8	**	**[‡]	4%	0%	9	4%	193
	ELA	9	**	**	5%	0%	14	4%	263
	ELA	10	**	**	7%	0%	14	5%	198
	ELA	11	**	**	4%	0%	9	3%	234
	Algebra 1	--	0%	74%[‡]	18%	0%	37	12%	172
	Algebra 2	--	0%	76%[‡]	9%	1%	18	6%	180
	Geometry	--	0%	76%	10%	0%	23	7%	201
	Int. Math 1	--	0%	55%	23%	0%	9	11%	53
	Int. Math 2	--	0%	68%	23%	0%	8	9%	49
	Int. Math 3	--	0%	64%	8%	0%	4	4%	49
	Math	3	0%	41%	4%	0%	10	3%	226
	Math	4	0%	40%	3%	0%	7	2%	238
	Math	5	0%	43%	4%	0%	9	3%	204
	Math	6	1%	47%	1%	0%	3	1%	203
	Math	7	0%	66%	6%	0%	13	5%	220
	Math	8	0%	62%	6%	0%	16	5%	236

Note. ELA = English Language Arts.
[‡] One item had no correct student responses after drops were performed and was excluded from further analyses. **Max scores for PBA ELA items were not correct in the test map files. Mean proportion correct could not be computed.

Table 2 shows that, across the assessments, content areas, and grade levels and test forms, between 0-1% of items were flagged for high p-values (greater than .80 for polytomous items or .95 for dichotomous items). On the other hand, between 18% (EOY math grade 4) and 76% (PBA Algebra 2 and Geometry) of items were flagged for having a p-value less than .30. Between 1% (PBA math grade 6) and 32% (EOY Integrated Math 2) of items were flagged for low item-total correlations. Few items across the assessments (between 0% and 4%) were flagged for having negative item-total correlations.

Table 2 indicates the final number of items designated suitable for further analyses based on item statistics and administrative drops. Some of these items were designated for a special vertical linking study and were given in “off-grade” conditions. Items were always administered to an adjacent grade (so grade 4 items were never administered to grade 6). It is unknown if these items can be repurposed for operational administration, but they were not included in the administrative drop list as DNU. These items are therefore included in subsequent analyses. If they cannot be used operationally, the number of potential on-grade items will be reduced substantially. Table 3 indicates the total number of items administered on each test designated for a lower grade, the administered grade (or on-grade level), and the higher grade. The total items column indicates the available non-dropped items and is repeated from Table 2. The remaining columns indicate the percentage of non-dropped items that are designated as "on-grade level" versus the total designated as "off-grade level." In some instances the proportion of off-grade level items approaches 30%. Some high school subject-level tests had no off-grade items (Algebra II, Integrated Math II and III).

Table 3. Items Designated for Vertical Scaling Study (Off-Grade Items)

			Designated Grade Level			Statistical Drops Identified by HumRRO		
Test	Subject	Grade	Lower Grade	On Grade Level	Higher Grade	Total Items	% On Grade Level	% Off Grade Level
EOY	ELA	3		207	24	231	89.6	10.4
	ELA	4	27	222	25	274	81.0	19.0
	ELA	5	32	182	21	235	77.4	22.6
	ELA	6	25	212	15	252	84.1	15.9
	ELA	7	16	196	19	231	84.8	15.2
	ELA	8	20	182	17	219	83.1	16.9
	ELA	9	15	206	15	236	87.3	12.7
	ELA	10	21	221	15	257	86.0	14.0
	ELA	11	20	252		272	92.6	7.4
	Algebra 1	--	39	287		326	88.0	12.0
	Geometry	--	59	306		365	83.8	16.2
	Int. Math 1		22	120		142	84.5	15.5
	Math	3		399	56	455	87.7	12.3
	Math	4	49	210	19	278	75.5	24.5
	Math	5	41	191	18	250	76.4	23.6
	Math	6	43	292	18	353	82.7	17.3
	Math	7	40	298	16	354	84.2	15.8
	Math	8	35	348	20	403	86.4	13.6
PBA	ELA	3		169	11	180	93.9	6.1
	ELA	4	26	184	18	228	80.7	19.3
	ELA	5	39	136	14	189	72.0	28.0
	ELA	6	35	175	14	224	78.1	21.9
	ELA	7	37	138	19	194	71.1	28.9
	ELA	8	38	139	16	193	72.0	28.0
	ELA	9	29	220	14	263	83.7	16.3
	ELA	10	34	152	12	198	76.8	23.2
	ELA	11	31	203		234	86.8	13.2
	Algebra 1	--	22	150		172	87.2	12.8
	Geometry	--	28	173		201	86.1	13.9
	Int. Math 1	--	7	46		53	86.8	13.2
	Math	3		194	32	226	85.8	14.2
	Math	4	23	199	16	238	83.6	16.4
	Math	5	23	172	9	204	84.3	15.7
	Math	6	22	168	13	203	82.8	17.2
	Math	7	22	189	9	220	85.9	14.1
	Math	8	25	201	10	236	85.2	14.8

Correlations with other measures. If the PARCC items and tasks measure the intended CCSS knowledge and skills, then it would be expected that PARCC scores correlate in expected ways with other measures of a similar content domain. For example, if PARCC math scores correlated highly with state math scores, this would be evidence of convergent validity. Findings that two assessments of different content areas (such as PARCC ELA and state math) do not correlate as highly would serve as evidence of discriminant validity. Tables 4 through 6 present convergent and discriminant validity correlations for the states for which state assessment scores were obtained. Results are presented separately for each state to account for differences in reporting scales and possible differences in the specific content or structure of the state assessments.

Most of the convergent validity coefficients for grades 3-8 indicate moderately high convergence between PARCC and state assessment results (correlations of ~.6 and higher). In the high school grades, however, many of the convergent validity coefficients indicated weaker relationships. The relatively weak convergent validity correlations for many of the assessments indicate instances where the association between PARCC scores and state scores are not as high as would be expected if the two assessments were measuring common content domains.

Table 4. Correlations Between PARCC Scores (Proportion Correct) and State Assessment Scores for Elementary School Grades

		PARCC ELA				PARCC Math			
		EOY		PBA		EOY		PBA	
		State Reading	State Math	State Reading	State Math	State Reading	State Math	State Reading	State Math
Grade 3	State 1	0.70 (n = 1687)	0.66 (n = 1687)	.73 (n = 1912)	.67 (n = 1912)	.71 (n = 1441)	.77 (n = 1441)	.65 (n = 1554)	.74 (n = 1555)
	State 2	NR	NR	NR	NR	NR	NR	NR	NR
	State 3	0.72 (n = 2331)	0.62 (n = 2327)	.69 (n = 2538)	.58 (n = 2533)	.67 (n = 3096)	.84 (n = 2765)	.67 (n = 2804)	.77 (n = 2331)
	State 4	.79 (n = 115)	.70 (n = 4918)	.80 (n = 113)	.68 (n = 4966)	.69 (n = 4518)	.67 (n = 293)	.68 (n=4537)	.66 (n=292)
	State 5	0.75 (n = 762)	0.71 (n = 766)	.75 (n = 938)	.70 (n = 937)	.71 (n = 806)	.81 (n = 809)	.70 (n = 802)	.77 (n = 807)
Grade 4	State 1	.72 (n = 1626)	.65 (n = 1626)	.73 (n = 2151)	.62 (n = 2151)	.67 (n = 1651)	.77 (n = 1653)	.61 (n = 1615)	.72 (n = 1616)
	State 2	.68 (n = 463)	.60 (n = 465)	.74 (n = 265)	.65 (n = 265)	.68 (n = 520)	.76 (n = 522)	.47 (n = 141)	.60 (n = 142)
	State 3	0.72 (n = 2457)	0.68 (n = 2454)	.76 (n = 2668)	.65 (n = 2663)	.66 (n = 2400)	.81 (n = 2309)	.63 (n = 2634)	.76 (n = 2258)
	State 4	.74 (n = 4885)	.71 (n = 5141)	.74 (n = 4945)	.72 (n = 5202)	.69 (n = 4773)	.72 (n = 4582)	.67 (n=4819)	.68 (n=4626)
	State 5	NR	NR	.69 (n = 714)	.64 (n = 719)	NR	NR	.58 (n = 571)	.68 (n = 572)
Grade 5	State 1	.62 (n = 1482)	.58 (n = 1482)	.68 (n = 2289)	.60 (n = 2292)	.58 (n = 1463)	.68 (n = 1464)	.53 (n = 1812)	.63 (n = 1814)
	State 2	.67 (n = 453)	.59 (n = 454)	.74 (n = 101)	.73 (n = 104)	.62 (n = 570)	.72 (n = 569)	.25 (n = 115)	.51 (n = 114)
	State 3	.73 (n = 2605)	.65 (n = 2604)	.74 (n = 2930)	.65 (n = 2920)	.62 (n = 2740)	.82 (n = 2551)	.62 (n = 2935)	.75 (n = 2468)
	State 4	.75 (n = 686)	.71 (n = 687)	NR	NR	.58 (n = 409)	.74 (n = 409)	NR	NR
	State 5	.70 (n = 616)	.71 (n = 618)	.69 (n = 1020)	.67 (n = 1026)	.53 (n = 878)	.72 (n = 886)	.51 (n=831)	.67 (n=837)

Notes. Yellow highlighting denotes convergent validity coefficients. State 4 includes 2013 state test scores for students who participated in PARCC (Grade 3 students had no 2013 state test data). All correlations are statistically significant (p < .001). NR= Not reported because n < 100.

Table 5. Correlations Between PARCC Scores (Proportion Correct) and State Assessment Scores for Middle School Grades

		PARCC ELA				PARCC Math			
		EOY		PBA		EOY		PBA	
		State Reading	**State Math**	**State Reading**	**State Math**	**State Reading**	**State Math**	**State Reading**	**State Math**
Grade 6	State1	.69 (n = 1413)	.65 (n = 1413)	.73 (n = 1837)	.68 (n = 1837)	.59 (n = 1512)	.76 (n = 1516)	.60 (n = 1560)	.73 (n = 1563)
	State 2	.60 (n = 130)	.57 (n = 130)	.69 (n= 114)	.57 (n= 114)	.51 (n = 277)	.56 (n = 281)	NR	NR
	State 3	.74 (n = 2264)	.65 (n = 2261)	.73 (n = 2689)	0.64 (n = 2689)	.66 (n = 2613)	.81 (n = 2511)	.66 (n = 2885)	.76 (n = 2514)
	State 4	.71 (n = 4198)	.67 (n = 4457)	.67 (n = 4249)	.63 (n = 4521)	.67 (n = 4017)	.74 (n = 3856)	.64 (n=4055)	.69 (n=3890)
	State 5	.59 (n = 696)	.60 (n = 696)	.63 (n = 1037)	.61 (n = 1037)	.60 (n = 581)	.74 (n = 581)	.56 (n = 659)	.73 (n = 662)
Grade 7	State1	.64 (n = 1383)	.64 (n = 1383)	.67 (n = 1932)	.64 (n = 1933)	.55 (n = 1297)	.72 (n = 1297)	.52 (n = 1354)	.71 (n = 1355)
	State 2	.59 (n = 259)	.52 (n = 261)	NR	NR	.38 (n = 127)	.60 (n = 129)	NR	NR
	State 3	.65 (n = 2396)	.63 (n = 2393)	.71 (n = 2949)	.66 (n = 2938)	.64 (n = 2395)	.81 (n = 2306)	.60 (n = 2684)	.78 (n = 2457)
	State 4	.72 (n = 4443)	.68 (n = 4631)	.69 (n = 4504)	.66 (n = 4693)	.65 (n = 4081)	.71 (n = 3953)	.63 (n=4136)	.69 (n=4001)
	State 5	.75 (n = 505)	.69 (n = 508)	.72 (n = 1061)	.69 (n = 1065)	.55 (n = 688)	.69 (n = 691)	.57 (n = 741)	.72 (n = 743)
Grade 8	State1	.58 (n = 1393)	.60 (n = 1395)	.67 (n = 1887)	.65 (n = 1891)	.44 (n = 1423)	.66 (n = 1423)	.44 (n = 1500)	.68 (n = 1500)
	State 2	.74 (n = 137)	.66 (n = 137)	NR	NR	.52 (n = 140)	.57 (n = 140)	NR	NR
	State 3	.68 (n = 2403)	.66 (n = 2403)	.66 (n = 2707)	.61 (n = 2701)	.58 (n = 2791)	.80 (n = 2745)	.57 (n = 2761)	.75 (n = 2594)
	State 4	.75 (n = 656)	.71 (n = 659)	NR	NR	.67 (n = 552)	.80 (n = 553)	NR	NR
	State 5	NR	NR	.65 (n = 681)	.60 (n = 680)	NR	NR	.53 (n = 495)	.63 (n = 495)

Notes. Yellow highlighting denotes convergent validity coefficients. State 4 includes 2013 state test scores for students who participated in PARCC. All correlations are statistically significant (p<.001). NR= Not reported because n < 100.

Table 6. Correlations Between PARCC Scores (Proportion Correct) and State Assessment Scores for High School Grades

		PARCC ELA						PARCC Math			
		EOY		PBA				EOY		PBA	
		State Reading	State Math	State Reading	State Math			State Reading	State Math	State Reading	State Math
Grade 9	State 1	NR	.55 (n = 1034)	NR	.59 (n = 1566)	Algebra 1	State 1	.18 (n = 192)	.32 (n = 1115)	.34 (n = 213)	.42 (n = 1308)
	State 2	.46 (n = 212)	.29 (n = 212)	NR	NR		State 2	.53 (n = 257)	.59 (n = 258)	NR	NR
	State 3	NR	NR	NR	NR		State 3	NR	NR	NR	NR
	State 4	.66 (n = 1721)	.60 (n = 1722)	.68 (n = 2775)	.61 (n = 2771)		State 4	NR	NR	NR	NR
	State 5	NR	NR	NR	NR		State 5	NR	NR	.22 (n = 448)	.41 (n = 450)
Grade 10	State 1	NR	.54 (n = 1026)	NR	.54 (n = 1023)	Algebra 2	State 1	.38 (n = 527)	NR	.36 (n = 796)	NR
	State 2	NR	NR	NR	NR		State 2	NR	NR	NR	NR
	State 3	.60 (n = 1799)	.54 (n = 1796)	.52 (n = 242)	.39 (n = 241)		State 3	.35 (n = 323)	.52 (n = 323)	NR	NR
	State 4	NR	NR	NR	NR		State 4	.36 (n = 169)	.54 (n = 169)	.41 (n = 270)	.52 (n = 270)
	State 5	NR	NR	NR	NR		State 5	.27 (n = 285)	.35 (n = 307)	.36 (n = 515)	.42 (n = 580)
Grade 11	State 1	.53 (n = 976)	NR	.66 (n = 1437)	NR	Geometry	State 1	.38 (n = 179)	.62 (n = 982)	.38 (n = 194)	.56 (n = 1231)
	State 2	NR	NR	NR	NR		State 2	NR	NR	NR	NR
	State 3	NR	NR	NR	NR		State 3	.30 (n = 491)	.47 (n = 491)	.25 (n = 139)	.31 (n = 138)
	State 4	NR	NR	NR	NR		State 4	.44 (n = 672)	.61 (n = 672)	.34 (n = 616)	.51 (n = 615)
	State 5	NR	NR	.51 (n = 439)	.41 (n = 442)		State 5	NR	NR	NR	NR

Notes. Yellow highlighting denotes convergent validity coefficients. State 4 includes 2013 state test scores for students who participated in PARCC. All correlations are statistically significant (p<.001). NR= Not reported because n < 100. Correlations for Mathematics 1 and Mathematics 2 assessments for both EOY and PBA are not reported because all cells contained fewer than 100 students. Only State 3 had a sufficient number of students to report a correlation for EOY Mathematics 3 (ELA= .49; MA= .55).

Findings for Research Question 2: Do students interact with the items/tasks and stimuli as intended given the instructions?

Several sets of analyses were conducted to address research question 2. First, data from student and TA surveys were analyzed to determine if students and TAs reported that students understood the assessment directions and had opportunities to gain familiarity with the assessment format and item types. Next, patterns of item responses were analyzed to determine if students were more likely to omit responses to particular item types or to items measuring particular sub-claims. Next, the percentages of hand-scored responses that were identified as "off-topic" were calculated. Next, student behavior on multiple select items was analyzed to determine if students interacted with these item types in expected ways. Finally, person fit statistics were analyzed to determine the number of participating students identified as having a poor fitting pattern of item responses, and if these aberrant response patterns were more common among particular student subgroups.

Student and TA surveys. Several items from the student and TA surveys were related to student understanding of the instructions provided along with of the assessments. This section summarizes responses to the relevant survey items.

Student responses. Two questions from the student surveys measured students' perceptions about the accessibility of both the overall test directions and item-level instructions. Table 7 indicates that the majority of students across all tests and testing formats who provided responses indicated that they did understand all of the directions read to them by the test administrator. However, over 20% of students who participated in the math EOY (computer- or paper-based) indicated that they did not understand all of the directions.

Table 7. Student Responses: Did you Understand All of the Directions Read by the Person Who Gave You the Test?

	Yes		No	
Assessment	**n**	**%**	**n**	**%**
ELA				
Computer-based PBA	207,548	93.9	13,440	6.1
Computer-based EOY	138,008	85.9	22,722	14.1
Paper-based PBA	62,364	95.3	3,071	4.7
Paper-based EOY	48,302	90.8	4,871	9.2
Math				
Computer-based PBA	158,980	86.9	23,925	13.1
Computer-based EOY	122,948	78.0	34,765	22.0
Paper-based PBA	62,056	87.2	9,143	12.8
Paper-based EOY	47,011	78.5	12,886	21.5

Table 8 summarizes responses to the question asking students about how often they found the item-level directions hard to understand. The largest proportion of student respondents across the tests and testing formats indicated that the directions for questions on the test were hard to understand some of the time. Comparing the percentages of respondents who indicated that directions were difficult to understand most of the time or almost always across content areas suggests that students found math item directions more difficult than ELA item directions. Roughly one-quarter to one-third of student respondents indicated that directions provided for ELA questions (computer-based or paper-based) were almost never hard to understand.

Table 8. Student Responses: How Often was it Hard to Understand the Directions for the Questions on This Test?

	Almost Always		Most of the Time		Some of the Time		Almost Never	
Assessment	n	%	n	%	n	%	n	%
ELA								
CBT PBA	12,641	5.7	31,369	14.2	112,233	50.8	64,532	29.2
CBT EOY	11,088	6.9	24,542	15.3	80,066	49.9	44,832	27.9
PBT PBA	3,252	5.0	8,307	12.7	32,356	49.5	21,488	32.9
PBT EOY	2,838	5.3	6,693	12.6	25,273	47.6	18,344	34.5
Math								
CBT PBA	30,192	16.5	47,712	26.1	81,318	44.5	23,530	12.9
CBT EOY	21,824	13.9	34,926	22.2	72,208	45.8	28,555	18.1
PBT PBA	9,658	13.6	17,652	24.8	34,102	47.9	9,791	13.8
PBT EOY	6,842	11.4	1,2637	21.1	29,183	48.7	11,203	18.7

Students' ability to interact with the assessment items as intended could be influenced by opportunities to practice on similar types of items. Students participating in a computer-based assessment were asked an additional question regarding the number of times they had practiced on a computer or tablet to prepare for the PARCC assessment. Table 9 shows that over 40% of students across the assessments indicated that they had never practiced for the test, while approximately 35-38% indicated that they had practiced once. Approximately 55-57% of student respondents indicated that they had practiced at least once.

Table 9. Student Responses: How Many Times Did You Practice on a Computer or Tablet to Get Ready for This Test?

	Never		Once		More than once	
Assessment	n	%	n	%	n	%
ELA						
CBT PBA	97,735	44.4	78,325	35.6	44,102	20.0
CBT EOY	72,593	45.3	56,275	35.1	31,402	19.6
Math						
CBT PBA	80,286	44.1	68,584	37.6	33,311	18.3
CBT EOY	68,139	43.4	55,990	35.7	32,905	21.0

The final question on the student survey gave respondents an opportunity to provide an open-ended comment about any aspects of the test they found to be confusing or unclear. Responses were analyzed to identify major themes, several of which inform the question of whether or not students interacted with test items as intended. The themes below were mentioned in at least 5% of responses.

- Math PBA & EOY
  - Typing issues caused delays
    - Keyboards freezing; answers being erased; student's typing speed very slow
  - Trying to use equation symbols/calculator/other tools caused delays
  - Skipped questions to finish on time/finish early; did not show work due to lack of time
  - Directions for questions unclear/confusing; had to re-read several times, which caused delays
- English PBA & EOY
  - Text box, highlighting, typing, tool issues
    - Open response took longer b/c of typing, deleting answers, etc.
  - Harder to focus on test; technology was distracting
  - Directions for questions very unclear/confusing; had to re-read several times, which caused delays
  - Passages difficult to read/scroll through; took more time than on paper

TA responses. Two items on the TA survey measured TAs' beliefs about the adequacy of instructions provided to students during the test. Table 10 indicates that approximately 73% of computer-based test administrators agreed or strongly agreed that students appeared to understand the instructions read during the test administration. Among paper-based test administrators, approximately 50% agreed or strongly agreed with this statement. TAs for computer-based and paper-based administrations indicated more similar levels of agreement with the statement, "the instructions I read to the students covered all of the information necessary to take the test." Approximately 65% of computer-based test administrators and 54% of paper-based test administrators agreed or strongly agreed with this statement.

Table 10. TA Responses to Survey Items Related to Instructions Given to Students

	Strongly disagree		Disagree		Agree		Strongly Agree	
Assessment	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Students appeared to understand the instructions I read to them during test administration.								
Computer-based	256	5.8	921	20.9	2968	67.3	265	6.0
Paper-based	275	16.6	555	33.5	729	44.0	99	6.0
The instructions I read to the students covered all of the information necessary to take the test.								
Computer-based	285	6.5	1274	29.0	3607	59.4	225	5.1
Paper-based	214	12.9	556	33.5	793	47.7	98	5.9

Two additional survey items asked TAs about students' preparedness for the assessments through both practice with items and participating in a tutorial. Table 11 indicates that larger proportions of computer-based test administrators indicated that their students had practiced with PARCC sample items and participated in the PARCC tutorial. Approximately three-fourths of all computer-based test administrators indicated that students had practiced with PARCC items, and that students had completed the PARCC tutorial prior to the test administration. Approximately 13% computer-based test administrators indicated that students had not practiced with PARCC items, and that students had not completed the PARCC tutorial prior to the test administration.

Table 11. Teacher Responses to Survey Items Related to Students' Preparation for the Assessments

	Yes		No		Not sure/Don't know	
Assessment	n	%	n	%	n	%
Did students in your session(s) practice with PARCC sample items prior to administration?						
Computer-based	3,281	77.3	561	13.2	404	9.5
Paper-based	411	25.6	823	51.2	373	23.2
Did the students in your session(s) complete the PARCC tutorial prior to administration?						
Computer-based	3,146	74.1	574	13.5	528	12.4
Paper-based	119	7.4	1,016	63.1	476	29.5

TAs were also asked to provide open-ended comments about students' questions about the test-taking instructions. Responses were analyzed to identify major themes, several of which inform the question of whether or not students interacted with test items as intended:

- Computer-based test administrators:
  - Students had questions about using/locating tools and why certain tools that were available in the tutorial weren't available in the actual test. Additionally, students who had not done the practice tutorial spent test time playing with tools to become accustomed to using them.
  - Students had difficulty with both log on and exiting procedures; the instructions were not clear on how to begin the test or how to properly exit the testing session.
  - Students were confused about the instructions for extended time and what to do when they were finished with the test.
  - The length/structure/vocabulary of the instructions, while not directly questioned by students, was viewed as very overwhelming and/or repetitive by TAs. Several TAs commented that information overload led to student inattentiveness and loss of focus, which may have contributed to some of the issues stated above.
- Paper-based test administrators:
  - Switching between the answer document and test document; students were unsure if they could write in one or both documents, where to show their work, and how to fill in responses using the provided response grid.
  - Students in the math sections asked questions about using mixed numbers versus decimals, as well as when they could use a calculator.

- Students in the ELA sections asked questions about using a highlighter and how many of the passages they were supposed to reference in a constructed response answer.
- Students were confused about directions being read to them, as the page numbers and instructions did not match at times.
- Though not specific to student questions, many comments indicated instructions were too long or repetitive, causing students to stop paying attention; this may have contributed to some of the questions students had regarding instructions.

Omitted responses. Omitted response rates to items were analyzed by both sub-claim and item type. Relatively high rates of omitted responses for a particular sub-claim would indicate that students may have had difficulty with that particular subset of content, whereas high rates of omitted responses for a particular item type would indicate that students may have had difficulty with the item format itself.

Omit rates by sub-claim. Using student response data, the rates at which item responses were omitted (i.e., missing) for items designated as measuring a particular sub-claim or sub-claims were calculated. Table 12 presents these results for the ELA PBA assessments. The *N* values indicate the number of items measuring that sub-claim. The *mean value* is the average percentage of omitted responses for that set of items. The *max* column presents the highest percentage of omitted responses for that set of items.

Table 12 indicates that across the grade levels of ELA PBA tests, items measuring writing sub-claims tended to have the largest rates of omitted responses. Mean omit rates ranged from 1.44 (Grade 7 Reading Vocabulary) to 21.88 (Grade 11 Written Expression and Writing Knowledge Language and Conventions).

Table 12. Descriptive Statistics of Omit Rates by Sub-claim for the English/Language Arts Performance-Based Assessment

Grade	Sub-claim	N	Mean	SD	MIN	MAX
3	Reading-RI	36	3.28	1	1.3	5.03
	Reading-RI:Writing-WE:Writing-WKL	10	5.54	1.05	4.04	7.32
	Reading-RL	83	3.62	2.58	0.33	7.81
	Reading-RL:Writing-WE:Writing-WKL	9	3.29	0.52	2.63	4.4
	Reading-RV	42	2.05	1.59	0.13	6.36
	Writing-WE:Writing-WKL	6	7.64	2.49	4.89	10.5
4	Reading-RI	52	2.01	0.6	1.09	3.43
	Reading-RI:Writing-WE:Writing-WKL	12	3.97	0.71	2.85	5.39
	Reading-RL	108	2.93	2.33	0.08	8.18
	Reading-RL:Writing-WE:Writing-WKL	15	2.24	0.6	1.26	3.3
	Reading-RV	38	1.62	1.32	0.2	5.78
	Writing-WE:Writing-WKL	15	7.26	2.73	3.81	13.71

(continued)

Table 12. Descriptive Statistics of Omit Rates by Sub-claim for the English/Language Arts Performance-Based Assessment (continued)

Grade	Sub-claim	N	Mean	SD	MIN	MAX
5	Reading-RI	47	2.1	1.17	1.13	9
	Reading-RI:Writing-WE:Writing-WKL	13	3.9	0.58	3.08	5.06
	Reading-RL	76	2.99	1.91	0.17	7.44
	Reading-RL:Writing-WE:Writing-WKL	10	2.66	0.72	1.79	3.99
	Reading-RV	39	2.03	1.2	0.39	4.82
	Writing-WE:Writing-WKL	15	5.98	0.95	3.2	7.31
6	Reading-RI	44	2.21	0.63	1.12	3.73
	Reading-RI:Writing-WE:Writing-WKL	10	4.52	0.73	3.58	5.75
	Reading-RL	95	3.35	2.8	0.07	17.87
	Reading-RL:Writing-WE:Writing-WKL	13	2.28	0.52	1.26	2.87
	Reading-RV	51	1.77	1.44	0.1	5.53
	Writing-WE:Writing-WKL	13	7.48	1.39	5.63	10.68
7	Reading-RI	37	2.63	0.93	1.43	5.11
	Reading-RI:Writing-WE:Writing-WKL	7	5.71	0.95	4.05	6.56
	Reading-RL	80	2.78	2.27	0.05	7.91
	Reading-RL:Writing-WE:Writing-WKL	14	2.45	0.79	1.3	4.36
	Reading-RV	46	1.44	1.23	0.05	4.82
	Writing-WE:Writing-WKL	11	7.54	1.23	5.65	9.7
8	Reading-RI	34	2.62	0.63	1.5	4.19
	Reading-RI:Writing-WE:Writing-WKL	8	6.13	0.89	5.24	7.79
	Reading-RL	83	3.1	1.76	0.24	5.61
	Reading-RL:Writing-WE:Writing-WKL	14	3.7	1.21	2.29	6.76
	Reading-RV	50	1.93	1.21	0.26	4.92
	Writing-WE:Writing-WKL	13	8.4	1.57	6.96	12.18
9	Reading-RI	80	3.91	1.48	0.73	7.56
	Reading-RI:Writing-WE:Writing-WKL	17	10.37	1.81	5.57	13.21
	Reading-RL	86	3.65	2.96	0.05	11.64
	Reading-RL:Writing-WE:Writing-WKL	16	6.64	3.24	3.98	16.09
	Reading-RV	61	3.05	1.96	0.1	7.51
	Writing-WE:Writing-WKL	15	14.14	3.06	9.42	20.1
10	Reading-RI	59	4.43	1.39	2.28	8.96
	Reading-RI:Writing-WE:Writing-WKL	14	15.96	1.94	11.84	19.03
	Reading-RL	61	4.72	2.81	0.15	8.85
	Reading-RL:Writing-WE:Writing-WKL	9	12.11	2.83	6.21	15.31
	Reading-RV	55	3.67	1.99	0.08	8.06
	Writing-WE:Writing-WKL	13	20.13	3.01	15.1	26.16
11	Reading-RI	67	4.15	1.64	1.74	9.63
	Reading-RI:Writing-WE:Writing-WKL	14	17.79	2.22	13.53	20.8
	Reading-RL	82	5.14	5.76	0.08	46.35
	Reading-RL:Writing-WE:Writing-WKL	11	10.47	1.44	7.81	12.66
	Reading-RV	54	4.31	6.24	0.08	46.53
	Writing-WE:Writing-WKL	15	21.88	2.23	17.77	25.26

Note. Cells with multiple sub-claims represent items with multiple parts that each correspond to a different sub-claim. Reading-RI = Reading Information; Reading-RL = Reading Literature; Reading-RV = Reading Vocabulary; Writing-WE = Writing Written Expression; Writing-WKL = Writing Knowledge Language and Conventions.

Table 13 presents the omit rates by sub-claim for the ELA EOY assessments. Items measuring Writing sub-claims were not included in the EOY assessments. The average omit rates among the ELA EOY tests tend to be lower than the PBA rates, ranging from 1.6% omitted responses (Grade 6 Reading Language) to 3.98% omitted responses (Grade 10 Reading Information). Across the grades, items measuring the Reading Information sub-claim tend to have the highest average omit rates.

Table 13. Descriptive Statistics of Omit Rates by Sub-claim for the English/Language Arts End-of-Year Assessment

Grade	Sub-claim	N	Mean	SD	MIN	MAX
3	Reading-RI	94	3.02	1.61	0.27	7.09
	Reading-RL	101	2.06	1.64	0.08	6.26
	Reading-RV	44	2.55	1.53	0.16	5.01
4	Reading-RI	109	2.44	1.21	0.29	4.81
	Reading-RL	115	1.86	1.52	0.05	4.99
	Reading-RV	56	1.68	1.34	0.06	4.56
5	Reading-RI	76	2.92	1.2	0.6	5.49
	Reading-RL	110	1.87	1.66	0.1	6.59
	Reading-RV	59	1.79	1.48	0.07	4.74
6	Reading-RI	115	2.61	1.45	0.14	6.66
	Reading-RL	75	1.6	1.45	0.07	5.68
	Reading-RV	69	1.77	1.29	0.07	4.73
7	Reading-RI	96	2.89	1.59	0.18	6.69
	Reading-RL	78	3.34	7.17	0.09	46.89
	Reading-RV	64	2.32	1.66	0.08	7.45
8	Reading-RI	104	3.06	1.03	0.45	5.06
	Reading-RL	58	3.05	7.23	0.1	55.8
	Reading-RV	64	2.18	1.38	0.1	4.55
9	Reading-RI	102	3.72	1.93	0.19	8.77
	Reading-RL	71	3.04	2.78	0.08	11.65
	Reading-RV	77	2.71	1.92	0.06	8.49
10	Reading-RI	118	3.98	2.37	0.22	10.15
	Reading-RL	70	3.55	3.26	0.05	10.89
	Reading-RV	75	3.46	2.42	0.09	9.62
11	Reading-RI	121	3.92	1.7	0.42	8.6
	Reading-RL	69	2.55	2.09	0.12	7.48
	Reading-RV	79	2.86	1.97	0.05	6.31

Note. Reading-RI = Reading Information; Reading-RL = Reading Literature; Reading-RV = Reading Vocabulary.

Table 14 presents these results for the Math PBA assessments. Sub-claims C and D (Evidence and Modeling, respectively) tended to have the highest omit rates, particularly in the higher grade levels. Average omit rates ranged from 1.05 (Grade 4 Fluency) to 29.27 (Integrated Math 3 Reasoning).

Table 14. Descriptive Statistics of Omit Rates by Sub-claim for the Mathematics Performance-Based Assessment

Subject/Grade	Sub-claim	N	Mean	SD	MIN	MAX
3	A	162	3.21	1.92	0.06	9.66
	C	34	6.65	2.79	2.13	16.4
	D	27	7.06	4.05	0.86	20.88
	E	9	2.62	1.8	0.15	5.2
4	A	165	4.43	2.86	0.12	12.07
	C	44	8.97	3.46	3.4	17.26
	D	33	8.59	3.54	3.24	16.55
	E	3	1.05	0.65	0.31	1.5
5	A	138	4.26	3.06	0.08	12.77
	C	39	8.21	2.85	2.55	15.63
	D	35	6.88	3.13	2.43	16
6	A	148	2.72	2.67	0.04	10.73
	C	29	10.43	3.05	2.89	17.39
	D	25	11.02	5.31	2.43	22.08
7	A	159	3.2	3.27	0.06	13.45
	C	36	14.74	6.26	3.93	32.65
	D	34	15.03	4.85	8.64	27.4
8	A	172	2.17	2.69	0.09	15.83
	B	3	3.19	0.83	2.25	3.8
	C	40	13.52	5.16	3.22	27.41
	D	36	14.5	5.56	5.75	24.36
Algebra 1	A	133	3.84	4.13	0.18	22.71
	B	20	2.42	1.83	0.28	8.45
	C	30	27.08	9.28	11.81	43.87
	D	25	23.6	10.29	11.07	47.77
Algebra 2	A	92	4.59	4.65	0.26	22.72
	B	58	3.62	4.27	0.07	20.16
	C	24	25.72	10.27	2.39	49.95
	D	24	23.39	10.8	2.07	45.76
Geometry	A	135	4.01	4.03	0.14	24.33
	B	21	1.04	0.9	0.15	3.39
	C	39	24.19	6.82	10.97	43.19
	D	28	26.99	8.16	14.58	45.67
Integrated Math 1	A	39	3.19	4.75	0.12	22.25
	B	5	1.48	0.89	0.08	2.14
	C	7	23.62	8.8	11.05	32.67
	D	11	24.1	10.89	6.17	43.3

(continued)

Table 14. Descriptive Statistics of Omit Rates by Sub-claim for the Mathematics Performance-Based Assessment (continued)

Subject/Grade	Sub-claim	N	Mean	SD	MIN	MAX
Integrated Math 2	A	36	2.97	3.5	0.27	15.24
	B	4	1.37	0.61	0.72	2.16
	C	4	25.38	6.61	15.83	30.58
	D	5	18.02	9.55	8.18	32.48
Integrated Math 3	A	41	6.33	6.9	0.46	33.18
	B	5	4.06	1.76	2.37	6.77
	C	4	29.27	10.63	19.07	43.95
	D	3	24.54	21.25	8.17	48.56

Note. Sub-claim A= Major content; Sub-claim B= Additional and supporting content; Sub-claim C= Reasoning; Sub-claim D= Modeling; Sub-claim E= Fluency.

Table 15 presents these results for the Math EOY assessments. Omit rates among items for which sub-claim information was available range from 1.61 (Grade 7 Fluency[6]) to 11.33 (Integrated Math 2 Additional and supporting content). Sub-claim E (Fluency) tends to have slightly higher omit rates among the lower grades. Among the higher grades, sub-claims A (Major content) and B (Additional and supporting content) tend to be the only sub-claims measured, and there is not a consistent pattern in which either has higher omit rates.

Table 15. Descriptive Statistics of Omit Rates by Sub-claim for the Mathematics End-of-Year Assessment

Grade	Sub-claim	N	Mean	SD	MIN	MAX
3		36	2.37	1.24	0.48	5.13
	A	275	2.65	1.67	0.04	6.99
	B	128	2.38	1.57	0.07	6.75
	E	38	4.02	1.56	0.69	8.6
4		181	3.12	2.23	0.04	10.12
	A	177	3.71	2.21	0.08	10.22
	B	77	3.7	2	0.27	7.94
	E	23	5.55	1.6	2.04	7.46
5		180	3.67	2.7	0.08	9.13
	A	162	3.75	2.83	0.08	14.51
	B	84	3.96	2.49	0.08	11.88
	E	11	7.5	2.25	3.45	11.18
6		13	3.48	2.2	0.23	7.29
	A	247	4.52	3.62	0.17	15.37
	B	95	4.02	3.14	0.17	14.33
	E	30	7.59	3.95	0.89	13.85

(continued)

[6] Fluency is not a measured construct for grade 7. This item may have been an off-grade item on a form for the linking study or may have been miscoded. There was only a single item coded as Fluency for grade 7.

Table 15. Descriptive Statistics of Omit Rates by Sub-claim for the Mathematics End-of-Year Assessment (continued)

Grade	Sub-claim	N	Mean	SD	MIN	MAX
7	A	281	3.15	2.48	0.15	12.03
	B	113	4.86	2.33	0.66	11.23
	E	1	1.61		1.61	1.61
8		2	3.96	0.93	3.31	4.62
	A	323	3.5	4.48	0.18	69.9
	B	105	3.34	2.02	0.11	8
Algebra 1		41	8.35	3.43	2.01	17.09
	A	219	4.67	3.21	0.1	15.76
	B	140	4.85	3.43	0.06	17.31
Algebra 2		47	8.77	3.32	1.39	16.86
	A	154	5.52	4.22	0.42	20.59
	B	142	6.53	4.27	0.43	22.83
Geometry	A	205	5.76	4.18	0.08	20.11
	B	194	6.06	4.58	0.07	18.15
Integrated Math 1		12	9.75	3.69	4.38	15.19
	A	104	6.62	3.84	0.18	17.72
	B	55	7.02	5.05	0.53	19.11
Integrated Math 2		2	21.61	0.29	21.4	21.82
	A	66	8.48	7	0.6	27.73
	B	49	11.33	7.34	0.39	22.91
Integrated Math 3		10	9.84	4.38	3.39	15.02
	A	72	6.91	5.93	0.27	30.71
	B	60	8.32	4.72	0.81	26.43

[a] Cells with no sub-claim represent items in a subject/grade with missing sub-claim information in the testmap.
Note. Sub-claim A= Major content; Sub-claim B= Additional and supporting content; Sub-claim C= Reasoning; Sub-claim D= Modeling; Sub-claim E= Fluency.

Omit rates by item type. Similar analyses were conducted to determine rates of omitted responses were higher among particular item types. Table 16 presents these results for the ELA PBA assessments. Mean omit rates range from 1.8 (Grade 6 paper-based multiple choice multiple selection items) to 17.31 (Grade 11 open-ended items). Though the open-ended item type is most frequently associated with the largest omit rate within a grade level, other item types were occasionally associated with similar or higher omit rates.

Table 16. Descriptive Statistics of Omit Rates by Item Type for the English/Language Arts Performance-Based Assessment

Grade	Item Type	N	Mean	SD	MIN	MAX
3	MC	2	7.61	0.26	7.43	7.80
	MX	146	3.00	2.10	0.13	7.81
	OE	25	5.23	2.17	2.63	10.50
	XI	13	3.97	2.25	1.11	7.53
4	MC	1	1.09	--	1.09	1.09
	MX	186	2.38	1.81	0.08	8.18
	OE	42	4.53	2.74	1.26	13.71
	XI	11	3.50	3.21	0.5	7.75
5	MC	2	1.55	0.07	1.5	1.59
	MM	1	1.72	--	1.72	1.72
	MX	149	2.44	1.47	0.17	6.4
	OE	38	4.39	1.58	1.79	7.31
	XI	10	3.66	3.11	0.48	9
6	MC	3	4.00	1.69	2.35	5.72
	MM	1	1.80		1.8	1.8
	MX	175	2.56	2.2	0.07	17.87
	OE	36	4.78	2.44	1.26	10.68
	XI	11	3.95	2.86	0.45	7.27
7	MX	157	2.31	1.78	0.05	6.94
	OE	33	4.82	2.52	1.3	9.7
	XI	5	4.20	3.51	0.37	7.91
8	MX	163	2.62	1.5	0.24	5.61
	OE	35	6.00	2.45	2.29	12.18
	XI	4	3.83	1.88	1.24	5.37
9	MC	3	5.71	1.64	3.85	6.98
	MX	206	3.50	2.19	0.05	8.98
	OE	48	10.30	4.06	3.98	20.1
	XI	18	4.22	3.1	0.41	11.64
10	MC	1	7.16	--	7.16	7.16
	MX	168	4.19	2.14	0.08	8.86
	OE	36	16.50	4.04	6.21	26.16
	XI	6	6.83	1.93	3.73	8.96
11	MX	185	4.68	5.1	0.08	46.53
	OE	40	17.31	5.03	7.81	25.26
	XI	18	3.68	3.11	0.17	9.6

Note. MC= Multiple choice items, either single selection (paper or computer) or multiple selection (computer only). MM=Multiple choice items with multiple selections (paper only). MX= Multi-part items. OE= Open-ended items. XI= Standalone, technology-enhanced items (computer only).

Table 17 presents these results for the ELA EOY assessments. Mean omit rates range from 1.29 (Grade 11 multiple choice items) to 5.84 (Grade 10 standalone technology-enhanced items). There appears to be no clear pattern in which a particular item exhibits higher rates of omitted student responses.

Table 17. Descriptive Statistics of Omit Rates by Item Type for the English/Language Arts End-of-Year Assessment

Grade	Item Type	N	Mean	SD	MIN	MAX
3	MC	1	5.18	--	5.18	5.18
	MM	2	3.13	0.73	2.62	3.65
	MX	211	2.38	1.58	0.08	6.18
	XI	25	3.62	1.93	0.47	7.09
4	MC	1	3.57	--	3.57	3.57
	MM	2	2.36	0.04	2.33	2.39
	MX	255	2.02	1.36	0.05	4.88
	XI	22	2.35	1.83	0.07	4.99
5	MC	3	2.57	0.45	2.26	3.08
	MX	228	2.09	1.54	0.07	6.59
	XI	14	3.43	1.62	0.62	5.49
6	MC	3	2.85	2.48	0.41	5.38
	MX	232	2.05	1.43	0.07	6.66
	XI	24	2.46	1.81	0.14	5.68
7	MX	219	2.88	4.47	0.08	46.89
	XI	19	2.90	1.81	0.38	6.91
8	MX	201	2.80	4.00	0.10	55.80
	XI	25	2.87	1.26	0.65	4.79
9	MX	233	3.18	2.19	0.06	11.65
	XI	17	3.63	2.77	0.14	10.17
10	MX	255	3.65	2.63	0.05	10.89
	XI	8	5.84	2.69	0.72	10.15
11	MC	3	1.29	0.15	1.18	1.46
	MX	253	3.22	1.92	0.05	7.48
	XI	13	4.51	2.70	0.62	8.60

Note. MC= Multiple choice items, either single selection (paper or computer) or multiple selection (computer only). MM=Multiple choice items with multiple selections (paper only). MX= Multi-part items. OE= Open-ended items. XI= Standalone, technology-enhanced items (computer only).

Table 18 presents the results for the math PBA assessments. Mean omit rates range from 1.23 (Integrated Math 2 multiple choice items) and 24.37 (Integrated Math 3 open-ended items). Across the grade levels, multi-part and open-ended items tend to have the highest mean omit rates.

Table 18. Descriptive Statistics of Omit Rates by Item Type for the Mathematics Performance-Based Assessment

Subject/Grade	Item Type	N	Mean	SD	MIN	MAX
3	GR	13	4.42	1.6	1.79	6.68
	MC	43	2.87	1.63	0.23	6.81
	MM	13	3.83	1.49	1.67	6.88
	MX	76	4.8	3.62	0.06	20.88
	OE	26	7.1	2.12	2.87	11.12
	XI	61	2.97	1.92	0.14	6.54
4	GR	15	5.05	1.88	2.61	8.67
	MC	47	4.48	2.87	0.17	11.65
	MM	11	4.47	2.98	1.34	10.82
	MX	90	7.06	4.11	0.25	17.26
	OE	29	8.22	3.23	3.4	15.07
	XI	53	3.82	2.77	0.12	10
5	GR	2	5.68	1.94	4.31	7.05
	MC	42	3.91	2.56	0.16	9.22
	MM	6	5.92	3.85	2.02	11.39
	MX	89	6.21	3.59	0.17	16
	OE	34	6.95	2.68	1.63	13.34
	XI	39	3.82	3.39	0.08	11.74
6	GR	8	6.42	1.52	4.32	8.47
	MC	42	2.07	1.92	0.09	7.62
	MM	9	3.68	2.37	0.86	9.16
	MX	63	6.97	5.46	0.11	22.08
	OE	29	9.32	4.08	2	17.39
	XI	51	1.95	2.57	0.04	10.72
7	GR	9	7.7	3.75	2.81	13.45
	MC	45	2.6	3.07	0.27	11.45
	MM	15	2.73	1.98	0.56	6.74
	MX	64	11.44	7.59	1.04	32.65
	OE	37	13	4.4	3.93	23.49
	XI	59	1.88	2.19	0.06	8.22
8	GR	10	9.6	3.35	4.76	15.83
	MC	67	1.69	1.99	0.18	9.83
	MM	14	1.89	0.72	0.87	3.36
	MX	69	9.9	8.59	0.09	27.41
	OE	36	11.62	3.7	5.75	20.9
	XI	55	1.98	1.99	0.09	9.12
Algebra 1	GR	9	10.58	5.55	4.97	22.71
	MC	81	2.27	2.07	0.19	10.8
	MM	8	3.39	2.56	0.92	7.88
	MX	52	20.5	12.87	0.59	47.77
	OE	21	22.52	9.12	11.55	45.09
	XI	37	3.14	2.51	0.18	8.68

(continued)

Table 18. Descriptive Statistics of Omit Rates by Item Type for the Mathematics Performance-Based Assessment (continued)

Subject/Grade	Item Type	N	Mean	SD	MIN	MAX
Algebra 2	GR	8	15.45	5.66	8.67	22.72
	MC	65	2.41	1.54	0.36	7.9
	MM	16	4.32	2	1.76	9.57
	MX	47	17.77	11.54	1.1	45.22
	OE	22	23.99	11.36	10.21	49.95
	XI	40	2.45	2.4	0.07	10.12
Geometry	GR	9	15.71	3.98	10.07	24.33
	MC	55	2.35	1.86	0.2	7.38
	MM	16	3.96	1.32	1.89	6.72
	MX	58	16.77	13.38	0.15	45.67
	OE	36	23.2	6.9	10.97	39.05
	XI	49	2.47	2.06	0.14	8.87
Integrated Math 1	GR	2	14.54	10.9	6.83	22.25
	MC	23	1.7	1.21	0.34	6.25
	MM	4	2.05	1.84	0.82	4.72
	MX	17	18.03	13.61	0.12	43.3
	OE	8	20.58	9.33	6.17	32.67
	XI	8	1.84	2.11	0.08	6
Integrated Math 2	GR	3	11.94	2.86	10.24	15.24
	MC	22	1.23	0.86	0.29	4.03
	MM	4	1.36	0.35	1.08	1.83
	MX	9	13.28	12.25	1.5	32.48
	OE	6	17.34	7.94	8.18	30.58
	XI	5	2.41	1.89	0.27	5.26
Integrated Math 3	GR	4	18.72	14.38	4.28	33.18
	MC	23	4.28	2.53	0.46	9.7
	MM	8	5.33	2.45	0.74	9.13
	MX	9	16.39	17.9	0.7	48.56
	OE	3	24.37	5.09	19.07	29.21
	XI	6	5.65	7.1	0.52	19.35

Note. MC= Multiple choice items, either single selection (paper or computer) or multiple selection (computer only). MM=Multiple choice items with multiple selections (paper only). MX= Multi-part items. OE= Open-ended items. XI= Standalone, technology-enhanced items (computer only). GR= Gridded response items (paper only).

Table 19 presents the results from math EOY assessments. Mean omit rates range from 1.97 (Grade 8 paper-based multiple choice multiple selection items) to 18.91 (Integrated Math 3 gridded response items). Overall, gridded response items tended to have higher mean omit rates across the grade levels.

Table 19. Descriptive Statistics of Omit Rates by Item Type for the Mathematics End-of-Year Assessment

Subject/Grade	Item Type	N	Mean	SD	MIN	MAX
3	GR	46	3.78	1.35	1.31	6.41
	MC	148	2.63	1.64	0.07	6.99
	MM	45	2.81	1.28	0.49	6.14
	MX	91	2.48	1.61	0.04	6.75
	XI	147	2.42	1.76	0.04	8.6
4	GR	50	5.07	2.08	1.33	10.22
	MC	156	3.38	2.04	0.08	7.31
	MM	26	3.09	1.92	0.52	7.94
	MX	118	3.41	2.21	0.04	7.93
	XI	108	3.43	2.37	0.08	10.12
5	GR	35	6.78	3.2	0.69	14.38
	MC	149	3.43	2.44	0.08	14.51
	MM	21	3.75	2.04	1.07	7.86
	MX	115	3.66	2.59	0.12	9.5
	XI	117	3.72	2.79	0.08	10.94
6	GR	51	7.15	2.91	2.57	14.49
	MC	108	2.75	2.48	0.3	10.96
	MM	18	2.48	1.35	1.36	6.4
	MX	96	5.8	3.2	0.48	14.33
	XI	112	4.53	4.23	0.17	15.37
7	GR	30	6.17	2.77	1.8	12.03
	MC	139	2.27	1.85	0.18	7.97
	MM	31	2.56	1.73	0.82	8.89
	MX	127	5.16	2.22	0.15	10.84
	XI	68	2.96	2.29	0.33	9.33
8	GR	30	6.82	3.06	2.65	13.46
	MC	155	2.45	2.01	0.11	8.38
	MM	35	1.97	1.44	0.54	6.64
	MX	125	4.45	6.25	0.26	69.9
	XI	85	3.27	2.15	0.18	7.26
Algebra 1	GR	9	8.77	2.96	3.54	12.44
	MC	132	2.92	1.82	0.1	8.43
	MM	28	3.93	1.59	1.83	6.92
	MX	180	6.83	3.8	0.35	17.31
	XI	51	4.68	2.66	0.06	10.59
Algebra 2	GR	10	15.48	3.33	8.94	20.59
	MC	107	3.23	1.41	0.42	6.54
	MM	16	3.66	1.27	1.29	5.76
	MX	177	8.2	4.02	0.71	22.83
	XI	33	5.47	3.53	1.18	12.25

(continued)

Table 19. Descriptive Statistics of Omit Rates by Item Type for the Mathematics End-of-Year Assessment (continued)

Subject/Grade	Item Type	N	Mean	SD	MIN	MAX
Geometry	GR	29	11.42	3.74	3.26	17.98
	MC	102	2.85	1.98	0.08	11.71
	MM	31	3.32	1.51	1.31	6.92
	MX	164	8	4.36	0.18	20.11
	XI	73	4.37	3.16	0.07	12.42
Integrated Math 1	GR	9	14.82	4.18	6.19	19.11
	MC	51	5.49	3.13	1.18	16.46
	MM	9	6.13	3.14	1.52	10.26
	MX	81	7.54	4.01	0.89	17.54
	XI	21	5.35	4.46	0.18	14.37
Integrated Math 2	GR	6	12.07	3.92	6.04	18.01
	MC	38	6.16	5.66	0.39	22.79
	MM	9	5.42	3.53	0.78	11.03
	MX	51	13.65	8.16	0.6	27.73
	XI	13	8.24	2.88	2.19	12.57
Integrated Math 3	GR	4	18.91	8.05	13.08	30.71
	MC	44	5.19	3.22	0.27	13.08
	MM	7	6.72	3.75	1.69	13.85
	MX	68	9.15	5.78	0.27	29.05
	XI	19	6.45	3.2	1.59	12.8

Note. MC= Multiple choice items, either single selection (paper or computer) or multiple selection (computer only). MM=Multiple choice items with multiple selections (paper only). MX= Multi-part items. OE= Open-ended items. XI= Standalone, technology-enhanced items (computer only). GR= Gridded response items (paper only).

"Off-topic" responses. A portion of items from the ELA PBA assessment were scored by hand, and the hand scoring process included procedures for identifying item responses that were deemed unrelated to the item, task, or stimulus. Scorers were trained to notify the scoring supervisor or scoring director if they believed a particular item response should be coded as "off-topic." This condition code could only be assigned by supervisory staff. Table 20 indicates that, on average, between 0.3% and 2.0% of items across the grade levels were rated as "off-topic."

Table 20. Percentages of "Off-Topic" Item Responses Among Hand-scored ELA PBA Items

Grade	Mean	Standard Deviation	Minimum	Maximum	Number of items
3	0.56	0.4	0.07	1.51	22
4	0.38	0.31	0.07	1.65	39
5	0.34	0.25	0.04	1.19	34
6	0.42	0.5	0.04	2.4	31
7	0.37	0.35	0.03	1.66	31
8	0.65	0.72	0.17	4.32	34
9	1.01	0.63	0.17	3.39	47
10	1.98	0.94	0.56	4.66	36
11	1.57	1.24	0.11	5.66	40

Student behavior on multiple select items. Several PARCC items were designed to collect multiple responses or pieces of evidence from students. These items were accompanied by instructions on the specific number of responses students were expected to provide. For example, a student might be asked to provide a response to a multiple choice question in the first portion of the item, and then identify multiple piece of evidence in the second portion of that item. Analyzing the pattern of responses provided thus offers additional evidence about whether or not students interacted with the PARCC items as intended. Tables 21 and 22 present the percentages of students providing each number of responses for a sample of items from four Grade 5 and Grade 8 forms, respectively. The number of responses each question prompted students for is highlighted in purple.

Across the two grade levels, the largest percentages of students (63% -94%) tended to provide the number of responses indicated in the item instructions. For the large majority of this item sample however, some percentage of students (1% - 9%) provided more responses than instructed, suggesting that some students did not interact with items as intended, given the item instructions. This may be indicative of students trying to select all possible response options because they did not know the correct answer, or students thinking that an incorrect response option was a correct response, even though they would be exceeding the requested number of responses.

Only two grade levels are included for these analyses. The student data file was not designed for analyses of part-level data. The information to conduct these analyses was only included in a very lengthy “xml” field of the data set. This code, though occupying only a single cell, may fill 30-40 pages of text if printed in a legible font. The xml includes the scoring instructions for the item. This code was translated into scoring instructions that could be applied to the student data file to generate the part data for these analyses. The process was time consuming and challenging, so we limited our analyses to only a sample of two grades.

Table 21. Percentages of Number of Student Responses to a Sample of Grade 5 Multiple Select Items

			Number of Responses Provided							
Form	Grade	Item.Part	0	1	2	3	4	5	6	7
014EO	5	2.02	0.15	34.14	62.96	2.17	0.31	0.15	0.10	0.15
014EO	5	7.02	0.57	27.58	68.80	2.48	0.57			0.57
014EO	5	9.01	0.83	24.38	72.42	1.86	0.36	0.15		0.83
014EO	5	13.01	1.50	22.06	74.54	1.45	0.21	0.26		1.50
014EO	5	13.02	1.70	20.25	75.83	1.86	0.10	0.15	0.10	1.70
014EO	5	14.02	4.08	20.45	73.81	1.39	0.21	0.05		4.08
014EO	5	17.02	2.32	26.14	69.42	1.70	0.26	0.15		2.32
014EO	5	24.02	5.11	18.75	73.45	2.22	0.21	0.26		5.11
014EO	5	25.02	5.22	18.49	74.64	1.34		0.26	0.05	5.22

(continued)

Table 21. Percentages of Number of Student Responses to a Sample of Grade 5 Multiple Select Items (continued)

			Number of Responses Provided							
Form	Grade	Item.Part	0	1	2	3	4	5	6	7
014EO	5	27.01	5.22	16.12	76.08	2.01	0.21	0.31	0.05	5.22
014EO	5	27.02	5.84	17.36	75.21	1.14	0.15	0.31		5.84
124EO	5	5.02	0.14	23.50	73.57	2.02	0.28	0.35	0.14	0.14
124EO	5	7.01	0.07	17.02	80.40	2.30	0.07	0.07	0.07	0.07
124EO	5	13.02	1.32	18.34	77.55	1.88	0.35	0.56		1.32
124EO	5	17.02	1.39	21.27	74.83	2.09	0.14	0.21	0.07	1.39
124EO	5	24.01	3.35	9.83	83.33	2.65	0.42	0.42		3.35
124EO	5	24.02	4.04	11.51	82.36	1.39	0.28	0.42		4.04
124EO	5	32.02	4.25	18.76	74.69	1.12	0.42	0.63	0.14	4.25
124EO	5	33.02	7.53	3.14	89.33					7.53
124EO	5	34.02	5.51	1.81	92.68					5.51
054EP	5	8.01	1.08	17.77	80.82	0.22	0.11			
054EP	5	8.02	1.19	18.42	79.96	0.43				
054EP	5	11.02	1.95	20.48	77.36	0.22				
054EP	5	16.01	3.03	17.77	77.68	1.52				
054EP	5	16.02	3.47	19.07	76.06	1.30	0.11			
054EP	5	20.01	1.52	15.17	82.77	0.54				
054EP	5	20.02	2.17	15.82	81.91	0.11				
054EP	5	22.02	1.84	17.77	80.39					
054EP	5	25.02	1.95	18.96	78.87	0.22				
054EP*	5	27.02	2.71	12.46	7.91	76.49	0.33	0.11		
054EP	5	30.02	2.93	19.07	77.57	0.43				
134EO	5	9.02	0.21	17.40	77.56	2.90	1.10	0.21	0.62	0.21
134EO	5	12.02	0.48	17.06	78.45	2.35	0.90	0.14	0.62	0.48
134EO	5	17.02	0.76	16.09	79.49	2.28	0.62	0.21	0.55	0.76
134EO	5	19.02	2.35	13.33	81.08	2.35	0.21	0.21	0.48	2.35
134EO	5	25.01	2.28	12.29	81.56	2.83	0.35		0.69	2.28
134EO	5	25.02	2.69	15.61	78.66	1.86	0.35	0.07	0.76	2.69
134EO	5	30.02	2.76	11.88	82.73	1.73	0.41	0.07	0.41	2.76

*Students were prompted to provide 3 responses.

Table 22. Percentages of Number of Student Responses to a Sample of Grade 8 Multiple Select Items

			Number of Responses Provided							
Form	Grade	Item.Part	0	1	2	3	4	5	6	7
024EO	8	1.01	10.18	6.52	83.30					
024EO*	8	2.02	0.51	20.16	3.36	73.83	1.83	0.20		0.10
024EO	8	5.02	0.31	24.03	69.96	4.79	0.81		0.10	
024EO*	8	12.02	0.31	17.31	3.26	77.80	1.22	0.10		
024EO*	8	18.02	3.67	9.57	1.93	82.99	1.63	0.20		
024EO	8	28.02	4.38	13.34	75.46	5.70	0.92		0.20	
024EO	8	34.02	4.58	11.10	80.65	2.95	0.61		0.10	
024EP	8	4.01	1.82	34.96	62.66	0.57				
024EP*	8	4.02	1.82	28.72	4.54	64.93				
024EP*	8	6.02	1.82	23.61	1.36	73.10	0.11			
024EP	8	7.02	2.50	24.40	70.83	2.16	0.11			
024EP	8	8.02	2.72	24.29	71.06	1.93				
024EP	8	11.02	3.41	22.93	72.19	1.48				
024EP*	8	12.02	3.63	22.25	3.63	70.37	0.11			
024EP	8	13.02	3.52	23.04	70.83	2.61				
024EP	8	23.02	4.43	18.84	76.05	0.68				
024EP	8	25.02	4.09	18.62	76.96	0.34				
024EP	8	27.02	4.88	20.20	74.57	0.34				
064EO	8	7.02	0.20	21.43	75.96	1.71	0.40		0.30	
064EO	8	9.02	0.40	12.37	79.28	7.14	0.30	0.10	0.40	
064EO	8	16.02	3.22	11.37	81.79	3.02	0.10	0.20	0.30	
064EO*	8	17.02	3.42	8.45	9.26	77.67	0.70		0.10	0.40
064EO	8	18.02	4.93	0.70	94.37					
064EO	8	23.02	3.32	7.04	86.72	2.21	0.20		0.50	
064EO	8	24.02	3.62	9.76	84.51	1.21	0.30	0.10	0.50	
064EO*	8	36.02	3.92	6.74	7.34	79.88	1.11	0.40	0.10	0.50
094EO	8	2.01	0.10	11.92	85.52	1.87	0.39		0.20	
094EO*	8	6.02	0.49	8.67	3.45	86.21	0.79	0.20	0.10	0.10
094EO	8	7.02	0.49	7.49	89.56	1.67	0.59	0.10	0.10	
094EO	8	9.02	0.39	7.00	82.36	9.75	0.39		0.10	
094EO	8	12.02	0.69	10.64	86.21	1.97	0.20	0.20	0.10	
094EO*	8	13.03	2.27	3.84	5.12	87.68	0.79	0.10		0.20
094EO*	8	20.02	4.04	6.50	2.86	84.93	1.28		0.10	0.30
094EO*	8	25.02	4.04	5.42	0.99	88.37	0.79			0.39
094EO	8	28.02	4.33	6.50	80.20	8.47	0.20		0.30	
094EO*	8	29.02	4.43	6.60	4.14	84.04	0.59			0.20
094EO	8	32.02	4.63	7.98	84.93	1.77	0.39		0.30	
094EO	8	33.02	4.63	7.59	85.32	1.97	0.20		0.30	
094EO	8	35.02	4.63	6.80	85.81	2.36	0.10		0.30	

Person fit. The purpose of a person fit analysis is to detect whether an examinee's responses aggregated across items are congruent with a specified IRT model (Reise, 1990), or are unlikely compared with the majority response pattern in the sample. Person fit indices are developed to investigate the consistency of a response pattern to the IRT model (Reise, 2000). The validity of individual test scores may be threatened when examinee's responses are governed by factors other than the ability of interest. Person fit analyses are used to detect response patterns to which external factors may be contributing, thus resulting in inaccurate or invalid measurement.

A sample of test forms were used for the person fit analyses, selected based on the overall number of students who were administered the form and to include a range of grade levels. The l_z person-fit statistic was used as it has been long considered one of most powerful and easily implemented statistics for the detection of non-fitting response patterns (Drasgow, Levine, and McLaughlin, 1991; Li & Olejnik, 1997) and has been widely applied and researched (Seo & Weiss, 2013).

Figure 2 depicts the distribution of the l_z person fit statistic for one Grade 7 Math EOY test form. The data points that fall below the dotted line have been flagged for poor person fit. The locations of the flagged cases along the X-axis indicate the ability level of the students. Figure 2 depicts few items flagged for poor person fit, a pattern that was observed across the sampled forms. Similar graphs for all the sampled forms are presented in Appendix B.

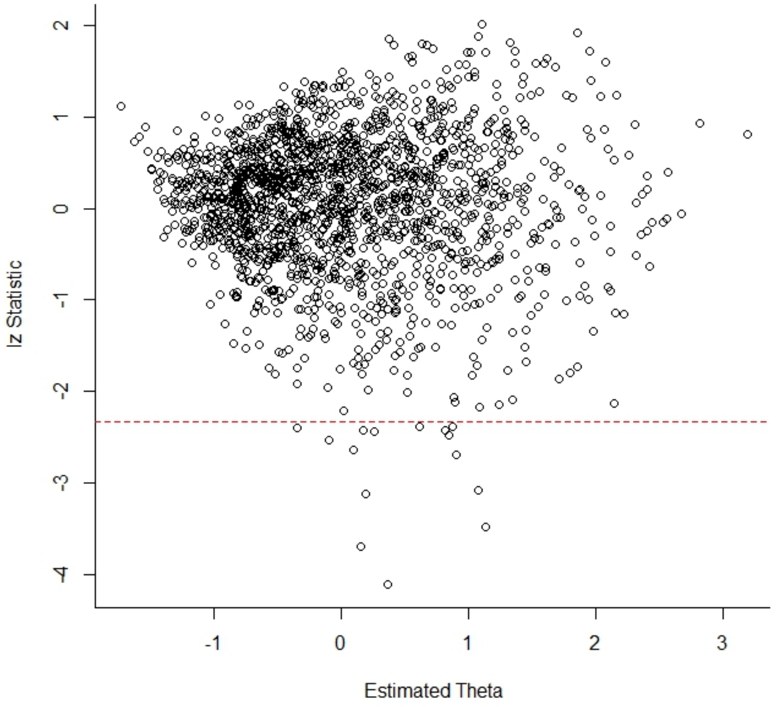


Figure 2. End-of-Year Math Grade 7 Form 104EO Person Fit Statistics

Looking at patterns of person fit statistics is useful for determining if particular student groups are demonstrating aberrant response patterns more frequently than other student groups. This would suggest that some group-related factors could be contributing to student performance on the assessment. Tables 23 and 24 present the distributions of the percentages of students identified as having a misfitting response pattern among gender, racial/ethnic, disability status, English language learner, and limited English proficiency subgroups for the PBA and EOY assessments.

The most noticeable pattern in both Table 23 and 24 is the relatively high percentage of students identifying with two or more ethnicities being flagged for poor person fit. This holds for both assessments with the exception of math PBA. However, this should be interpreted with caution given that there was only one form with an adequate sample of this subgroup to include in these analyses. Asians are also flagged more often than would be expected for mathematics. This may be due to a common pattern of Asian subgroups performing higher in mathematics than other subgroups combined with overall lower scores on PARCC.

Table 23. Performance-Based Assessment Person Fit Summary Information by Test Subject and Subgroup

	Performance-Based Assessment									
	Math					English/Language Arts				
	N Forms	Mean	SD	Min	Max	N Forms	Mean	SD	Min	Max
Sex										
Males	6	0.59%	0.37%	0.14%	1.12%	4	0.94%	0.31%	0.49%	1.20%
Females		0.95%	0.41%	0.51%	1.56%		0.91%	0.59%	0.10%	1.43%
Ethnicity										
White	6	0.78%	0.28%	0.44%	1.13%	4	0.98%	0.37%	0.44%	1.28%
Black		0.29%	0.46%	0.00%	1.17%		0.65%	0.25%	0.31%	0.89%
Hispanic		0.79%	0.55%	0.25%	1.46%		1.06%	0.75%	0.00%	1.68%
Asian	3	2.24%	2.94%	0.00%	6.56%	3	0.80%	0.70%	0.00%	1.25%
Two (+) Races	1	2.00%	0.00%	2.00%	2.00%	1	21.00%	0.00%	21.00%	21.00%
American Indian/Alaska Native	1	0.00%	0.00%	0.00%	0.00%		--	--	--	--
Disability Status										
Disability	6	0.79%	0.63%	0.00%	1.62%	4	0.96%	0.93%	0.00%	2.24%
No disability		0.74%	0.25%	0.41%	1.06%		0.92%	0.47%	0.26%	1.26%

(continued)

Table 23. Performance-Based Assessment Person Fit Summary Information by Test Subject and Subgroup (continued)

	Performance-Based Assessment									
	Math					English/Language Arts				
	N Forms	Mean	SD	Min	Max	N Forms	Mean	SD	Min	Max
ELL Status										
ELL	6	0.79%	1.18%	0.00%	3.03%	4	0.29%	0.57%	0.00%	1.15%
Non-ELL		0.78%	0.30%	0.49%	1.24%		0.95%	0.44%	0.30%	1.26%
LEP Status										
LEP	6	0.25%	0.40%	0.00%	0.91%	4	0.85%	1.07%	0.00%	2.22%
Non-LEP		0.80%	0.30%	0.52%	1.30%		0.94%	0.47%	0.25%	1.28%

Note. ELL= English Language Learner; LEP= Limited English Proficiency

Table 24. End of Year Assessment Person Fit Summary Information by Test Subject and Subgroup

	End of Year Assessment									
	Math					English/Language Arts				
	N Forms	Mean	SD	Min	Max	N Forms	Mean	SD	Min	Max
Sex										
Males	6	1.55%	1.37%	0.36%	4.11%	4	0.49%	0.28%	0.19%	0.78%
Females		1.82%	2.81%	0.00%	7.50%		0.26%	0.13%	0.11%	0.39%
Ethnicity										
White	6	1.94%	2.40%	0.26%	6.70%	4	0.48%	0.25%	0.20%	0.79%
Black		0.81%	0.75%	0.00%	1.89%		0.33%	0.31%	0.00%	0.67%
Hispanic		0.92%	1.70%	0.00%	4.38%		0.07%	0.13%	0.00%	0.26%
Asian	3	3.10%	1.25%	2.02%	4.48%	3	0.48%	0.84%	0.00%	1.45%
Two (+) Races	1	26.00%	0.00%	26.00%	26.00%	1	43.08%	0.00%	43.08%	43.08%
Disability Status										
Disability	6	0.84%	0.72%	0.00%	2.14%	4	0.35%	0.69%	0.00%	1.39%
No disability		1.76%	2.20%	0.21%	6.18%		0.40%	0.12%	0.29%	0.54%
ELL Status										
ELL	6	1.76%	2.81%	0.00%	6.38%	4	0.37%	0.75%	0.00%	1.49%
Non-ELL		1.67%	2.00%	0.19%	5.67%		0.39%	0.12%	0.28%	0.53%
LEP Status										
LEP	6	1.42%	1.61%	0.00%	3.33%	4	0.00%	0.00%	0.00%	0.00%
Non-LEP		1.65%	2.02%	0.18%	5.70%		0.41%	0.14%	0.28%	0.54%

Note. ELL= English Language Learner; LEP= Limited English Proficiency

Findings for Research Question 3: What is the optimal testing time required for individual or sets of items/tasks to allow students to demonstrate what they know and can do with regard to the target of measurement?

Two sets of analyses were conducted to address research question 3. First, results from the timing analysis portion of the study are summarized. Next, results from the analyses of the subset of items designated as math fluency items are summarized.

Timing analysis. This section summarizes HumRRO's investigation of the optimal time to allow for students to complete the PARCC assessments without creating speeded conditions while keeping the testing times as reasonable as possible. The analyses presented are based on timing data collected at the item level during the PARCC field test administration. Timing data were not available for item part. Because item functionality (e.g. select all, select one, drag and drop, etc.) is defined by item part, no time analyses at that level were completed. Timing data was aggregated by session and by form within grades to generate specific session time information. Unless otherwise indicated session time was used as the unit of analyses.

Indicated versus experienced session times. Field tests were administered in 2-3 sessions per test. Any breaks the students received during the tests were between sessions. Figure 3 contains a table copied from the field test administration manual and shows an example of how time was allotted for each testing session. The additional time indicated in the tables is not restricted for students with an "extra time" accommodation. Any student not completing the session during the session time is provided additional time, up to the amount listed in the tables. Students with the "extra time" accommodation may receive even more time to complete the session.

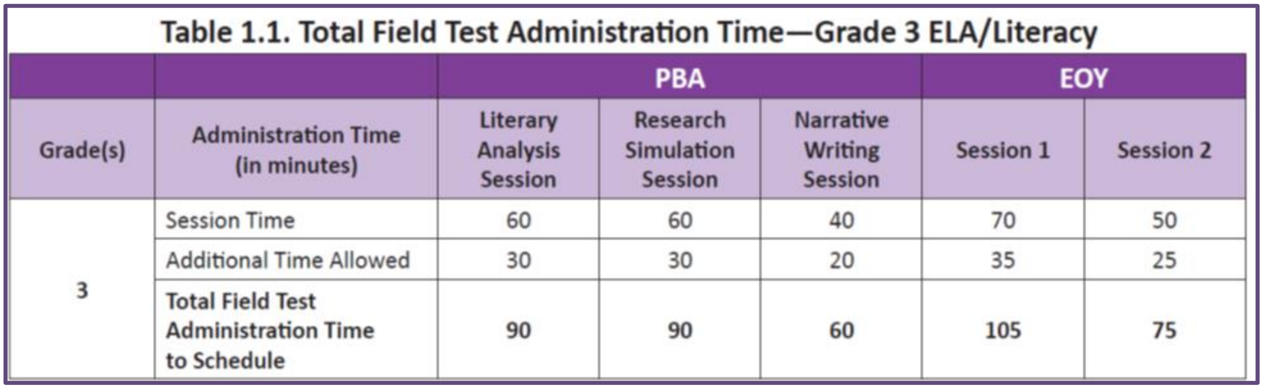


Table 1.1. Total Field Test Administration Time—Grade 3 ELA/Literacy

		PBA			EOY	
Grade(s)	Administration Time (in minutes)	Literary Analysis Session	Research Simulation Session	Narrative Writing Session	Session 1	Session 2
3	Session Time	60	60	40	70	50
	Additional Time Allowed	30	30	20	35	25
	Total Field Test Administration Time to Schedule	**90**	**90**	**60**	**105**	**75**

Figure 3. Excerpt of field test administration manual depicting allotted testing time for grade 3 ELA/Literacy.

If we look at the PBA assessment for grade 3 in Figure 3, which is administered in 3 sessions, we find that the first two sessions (Literary Analysis and Research Simulation) are each 60 minutes long, with 30 additional minutes that can be added if students need that time to finish. The

Narrative Writing session is shorter, at 40 minutes plus an additional 20 minutes if necessary. To investigate the reasonableness of the length of each session, it is informative to compare the time allotted for the sessions with the time students actually needed to complete the session.

Timing data were captured at the item level during the field test. The item time data was then aggregated to generate session timing data for each form. Paper versions of the test were not included in this analysis because no timing data were available for the paper tests. Students with obviously aberrant times were also omitted (e.g. students spending less than 1 minute total on the test). There were no students in the data file who had condition codes indicating no response for the last 5 items. Therefore, no students were omitted because they did not complete the test. It is possible that some students did not attend to the final items and gave random responses as time ran out, but we had no way to identify those students in the available data.

Analyses were completed at the session level. Table 25 provides an example of the results, showing the time students actually spent on the grade 3 ELA PBA. Similar tables for the other ELA grades and assessments are presented in Appendix C.

The first column of Table 25 indicates the session and the number of forms included in the analyses, as well as the initial session time and the total possible time allotted for each administration. When there are multiple numbers of forms indicated, this means that there were different numbers of forms for “no extra time” versus “extra time” indicated students. To be included in the analyses, forms must have at least 10 students. The second column indicates the amount of time in minutes the typical (median) student used to complete the session. The third column indicates the amount of time the typical (median) student with the “extra time” accommodation used to complete the session (Extra time accommodated results are italicized.). Forms were not weighted by numbers of students, so timing data from a form with 2000 students contributed the same to this analysis as a form with 40 students. This is appropriate since we are most interested in setting an optimal session time across forms. No sessions were omitted as outliers for the timing analyses. The last columns represent the time required for approximately 95% and 98% of students to complete the sessions. These estimates were computed assuming that the forms were roughly normally distributed. We used median plus 2 * the standard deviation for the 95% estimate, and median plus * 2.5 the standard deviation for the 98% estimate. Estimate were made by form and then aggregated.

Table 25. ELA PBA Timing Summary All Students (Grade 3 Only)*

Grade 3	Time (in minutes) for Typical (Median) Student		Time for About 95% Completion (+ 2 SD)		Time for About 98% Completion (+ 2.5 SD)	
Test, Number of forms, and Time Limits	**No Extra Time**	***Extra Time***	**No Extra Time**	***Extra Time***	**No Extra Time**	***Extra Time***
Literary Analysis (N = 22/21) 60 mins, 90 mins	36.1	*37.2*	71.7	*76.3*	80.7	*86.1*
Research Simulation (N = 22/21) 60 mins, 90 mins	42.0	*42.6*	79.9	*82.7*	89.4	*92.8*
Narrative Writing (N = 22/21) 40 mins,60 mins	16.7	*16.5*	42.5	*45.1*	48.9	*52.3*

*Forms omitted with fewer than 10 students.

Table 25 shows that the median grade 3 student completed each session with considerable time left over. The median student used roughly two thirds of the allotted time for Literary Analysis and Research Simulation and less than half the allotted time for Narrative Writing. The median student with the “extra time” accommodation had very similar time data as non-accommodated students.

The remaining columns in Table 25 show the median time plus 2 and 2.5 standard deviations, respectively (for “no extra time” and “extra time” students). These time estimates should indicate the time required for approximately 95% and 98% completion rates (time for this percentage of students to complete the session). Note that these times are much larger than the median time. This is because there was considerable variance in the amount of time students took to complete each session. For all sessions, some of the additional time would be required for 95% of the students to complete the assessment. Most of the additional time would be required to achieve a 98% completion rate.

A similar pattern was observed across the ELA assessments. The median time spent per session was typically one-half to two-thirds the allotted time. However, the large amounts of variance indicate that a number of students used the extra time made available to them.

Mathematics tests included grade-level tests for grades 3-8 for both PBA and EOY, as well as course-specific tests for algebra 1 and 2, geometry, and integrated mathematics 1, 2, and 3. Mathematics tests were all administered in 2 sessions. The use of additional time followed the same protocol as with the ELA assessments. Students with the “extra time” accommodation could receive even more time than is listed in the total administration time rows.

Table 26 presents a comparison of the actual time students needed to complete the test to the time allotted, using math PBA grades 3-5 as an example. This table follows the same format as the ELA tables above, and tables for all grades and tests are presented in Appendix C.

Table 26. Math PBA Timing Summary (Grades 3-5)

	Typical Student's Time (Median)		Time for About 95% Completion (+ 2 SD)		Time for About 98% Completion (+ 2.5 SD)	
	No Extra Time	*Extra Time*	No Extra Time	*Extra Time*	No Extra Time	*Extra Time*
Grade 3						
Session 1 (N = 14/12) 50 mins, 75 mins	41.1	42.5	74.8	82.3	83.2	92.3
Session 2 (N = 14/12) 50 mins, 75 mins	33.4	33.9	63.2	67.5	70.6	75.9
Grade 4						
Session 1 (N = 15) 50 mins, 75 mins	43.0	44.6	74.0	87.7	81.7	98.5
Session 2 (N = 15) 50 mins, 75 mins	35.5	35.9	63.8	76.8	70.9	87.0
Grade 5						
Session 1 (n=12) 50 mins, 75 mins	48.7	53.0	81.6	103.7	89.8	116.4
Session 2 (n=12) 50 mins, 75 mins	35.8	37.3	65.5	87.3	72.9	99.8

*Forms omitted with fewer than 10 students.

The results in Table 26 indicate that the typical (median) student took more than 40 minutes to complete Session 1 of the mathematics PBA assessments for grades 3-5. Session 2 was shorter, with the median student needing about 35 minutes to complete it. Grades 4 and 5 are especially troubling, given that the median student used nearly the full amount of session time to complete the test. The data in the latter columns shows that, for most grades, the full additional time would be required for 95% of students to complete the test and additional time beyond that would be required for a 98% completion rate.

The EOY assessment median completion times seem more reasonable for grades 3-5. The median times for EOY also represent a larger proportion of the session times for math than for ELA. The median student completed the EOY after using roughly three-fourths of the session time. Nearly all of the additional time would be necessary to reach a completion rate of 98%.

This pattern of more students requiring at least the allotted time to complete the PBA assessment was observed across grade levels. Also, As with ELA, students receiving the "extra time" accommodation in grades did not exhibit dramatically longer completion times than students not receiving the accommodation.

Total students completing sessions within session time. All of the data in the preceding tables were based on sessions. So, it is possible that there were forms that did not adhere to the general conclusions drawn from looking at session as the unit of measure. This could lead to spurious conclusions if the forms were not spiraled evenly. For example, if the timing data were very different for a form that held most of the students, we could draw erroneous conclusions. To ensure that the conclusions above hold for the overall population of students taking the field test, we conducted an analysis of the percentage of students overall who completed each session within the session time (no additional time required). For a non-speeded test, we would expect 90% or more students to complete within the session time, without needing additional time. Table 27 presents these results.

Table 27. Percentage of Students Completing Each Session Within Session Time

ELA	EOY			PBA			
Grade	N	Session 1	Session 2	N	Session 1	Session 2	Session 3
3	23142	93.02%	91.72%	28327	87.28%	79.17%	92.29%
4	22796	93.29%	92.09%	30347	92.00%	93.07%	86.00%
5	20724	93.99%	94.69%	27280	90.75%	92.18%	87.00%
6	21506	97.63%	97.21%	27861	94.57%	92.71%	88.72%
7	20281	98.83%	98.78%	27433	95.36%	94.79%	93.41%
8	16860	99.27%	99.13%	26009	97.05%	95.06%	94.86%
9	16400	99.16%	98.94%	24147	96.91%	95.83%	94.79%
10	13433	99.02%	98.76%	19284	97.58%	96.59%	95.30%
11	7542	99.24%	99.59%	18187	98.26%	97.58%	97.09%
Math	EOY			PBA			
Grade	N	Session 1	Session 2	N	Session 1	Session 2	
3	15155	77.12%	87.58%	15155	68.91%	82.23%	
4	12913	74.98%	88.42%	12913	65.00%	82.33%	
5	17193	61.29%	86.86%	17193	50.47%	80.64%	
6	15668	76.52%	92.58%	15668	57.30%	82.32%	
7	13499	70.92%	89.47%	13499	49.53%	77.01%	
8	12202	86.28%	94.70%	12202	70.23%	86.17%	
	EOY			PBA			
	N	Session 1	Session 2	N	Session 1	Session 2	
Math I	1515	86.60%	98.88%	1010	68.22%	88.32%	
ALG I	13272	93.41%	98.90%	7070	59.45%	92.97%	
GEO	11866	94.57%	99.11%	6410	75.26%	92.68%	

(continued)

Table 27. Percentage of Students Completing Each Session Within Session Time (continued)

ELA	EOY			PBA			
	EOY			PBA			
	N	Session 1	Session 2	N	Session 1	Session 2	
Math II	971	94.03%	99.69%	471	77.49%	96.60%	
	EOY			PBA			
	N	Session 1	Session 2	N	Session 1	Session 2	
Math III	843	88.73%	99.76%	514	63.23%	95.53%	
ALG II	11886	92.01%	99.14%	7993	80.55%	95.45%	

Table 27 shows that ELA EOY session times are generous. In fact, session times could be reduced for grades 6-11 and more than 90% would still complete each session within the session time. The percentage completing each session is very similar across sessions within ELA EOY as well. The ELA PBA is more complicated to interpret. Time is generous for all sessions for grades 7-11 for the PBA. Time for Session 2, Research Simulation, may not have been sufficient for Grade 3 without the additional time. Grade 3 Session 1, Literary Analysis, was questionable regarding sufficient time for completion. Most students in grade 3 were able to complete Session 3, Narrative Writing in the allotted time, but many students did not finish Session 3 in grades 4-6.

Table 27 also shows that many students taking math PBA, especially Session 1, did not finish within the session time. Fewer than half completed Session 1 of the grade 7 PBA during the session time and other subject tests also had many students that did not complete the PBA within the session time.

The timing data for EOY Session 1 was within our expectations. The session times seem generally reasonable given this data, except for Math I and Math III, which did not reach our 90% threshold (the sample sizes for the integrated math tests were fairly small). The time allowed for EOY Session 2 was generous for all high school math subjects.

Interquartile ranges. Perhaps the most compelling timing data comes from the examination of the interquartile ranges. The center of the student time distribution provides a clear indication of the time requirements to complete the PARCC assessments for typical students. These data were aggregated across forms and is presented by session in Tables 28 and 29. The interquartile range represents the center 50% of the data. In this case, we present the time at the 25th percentile (at this time 25% of students have completed the session) and the 75th percentile (at this time 75% of students have completed the session). Data are bolded and italicized where students' interquartile range fell outside the session time. These test sessions are especially concerning given that many of the “center distribution” students did not finish in the allotted time.

Table 28. PBA Interquartile Ranges of Timing Data

Grade	session 1_25	session 1_75	Session Time 1	session 2_25	session 2_75	Session Time 2	session 3_25	session 3_75	Session Time 3
ELA									
3	26.60	49.83	60	31.95	56.98	60	9.92	27.05	40
4	28.88	52.43	70	31.23	59.25	80	19.63	43.02	50
5	29.72	53.53	70	32.58	60.37	80	19.70	41.68	50
6	30.38	55.48	80	32.18	59.88	80	19.23	40.30	50
7	28.83	52.72	80	28.60	54.35	80	16.60	34.58	50
8	27.22	48.83	80	28.68	54.60	80	15.37	32.18	50
9	25.88	48.53	80	24.43	50.58	80	14.58	31.82	50
10	20.12	43.79	80	21.20	48.37	80	10.93	29.88	50
11	20.72	42.53	80	19.05	44.52	80	11.15	27.87	50
Math									
3	***31.70***	***53.23***	***50***	24.98	44.70	50			
4	***33.33***	***54.78***	***50***	26.35	45.42	50			
5	***38.80***	***61.93***	***50***	27.13	46.70	50			
6	***37.17***	***58.80***	***50***	25.73	45.50	50			
7	***40.72***	***62.32***	***50***	28.13	48.78	50			
8	***33.23***	***52.17***	***50***	22.98	42.62	50			
Algebra 1									
	40.27	***63.70***	***55***	19.53	38.97	55			
Geometry									
	34.13	54.83	55	20.27	40.42	55			
Algebra 2									
	38.66	61.60	65	21.53	43.58	65			
Math I									
	36.28	***58.17***	***55***	22.72	44.48	55			
Math II									
	36.70	58.53	60	21.97	40.27	60			
Math III									
	46.83	***72.43***	***65***	25.22	43.60	65			

Table 29. EOY Interquartile Ranges of Timing Data

Grade	session1_25	session1_75	Session Time 1	session2_25	session2_75	Session Time 2
ELA						
3	30.92	52.25	70	21.20	38.23	50
4	31.42	51.95	70	21.55	38.25	50
5	30.75	51.13	70	20.17	35.15	50
6	26.15	44.02	70	24.73	43.90	70
7	23.38	38.77	70	22.45	39.57	70
8	21.41	35.50	70	19.85	37.48	70
9	19.32	34.67	70	16.88	36.32	70
10	18.43	33.92	70	15.72	35.93	70
11	17.78	34.58	70	12.12	32.30	70
12	17.76	36.45	70	8.33	31.34	70
Math						
3	31.70	53.23	55	24.98	44.70	55
4	33.33	54.78	55	26.35	45.42	55
5	***38.80***	***61.93***	***55***	27.13	46.70	55
6	37.17	58.80	60	25.73	45.50	60
7	***40.72***	***62.32***	***60***	28.13	48.78	60
8	33.23	52.17	60	22.98	42.62	60
Algebra 1						
	26.97	51.45	70	13.15	33.47	70
Geometry						
	28.05	50.87	70	12.20	34.23	70
Algebra 2						
	24.02	53.43		9.67	31.03	
Math I						
	27.16	56.98	70	12.61	33.18	70
Math II						
	17.68	47.70	70	7.78	25.47	70
Math III						
	24.37	54.88	70	8.85	28.15	70

The full computed percentile data across forms is presented in Appendix D. It includes the data above, as well as the breadth of the interquartile ranges and the 95th and 98th percentile data.

Survey data. Student and administrator surveys included items regarding timing of the PARCC assessments. These data can be used to help verify the findings from the timing data analyses. Administrator data support the findings described above. Student questionnaire data are less easily interpreted and it is not clear that students knew when they used additional time versus the allotted session time. This data can be found in Appendices E and F.

Administrators did provide open-ended comments related to timing. Themes include:

- Many ELA administrators stated that there was too much time for the literature and research sections, while some indicated that students needed more time for the narrative writing task.
- Younger students had more difficulty completing the English portion within the allotted time, as they were not proficient in typing (which was required for the prose and open-ended items).
- Many math administrators considered the items too difficult for the lower grade levels (3-5) and some had issues at the 6-8th grade level as well. Administrators indicated that students spent long periods of time on items they were not capable of correctly answering[7].
- Math item difficulty caused many students to become frustrated/feel rushed, so they completed the test very quickly.
- Conversely, other administrators noted that there wasn't enough time for the math because of item difficulty and technological issues (e.g., text boxes not working, tools not working, connectivity).

Math fluency analysis. The PARCC field test for grades 3 through 6 Mathematics included items identified as measuring mathematical fluency (MF). There were between 7 and 12 such items for each grade along with other Mathematics items for a total of up to 50 items per form. The MF items for each grade were embedded in three different test forms:

1. MF items were interspersed throughout one of the Math EOY forms.
2. MF items were at the end of a second EOY section, but no instructions indicating that speed was a factor.
3. MF items were at the end of a third EOY section, in a separately timed section, with special instructions indicating that speed counted.

It should be noted that the math fluency analyses presented here are based on a data set that was produced before corrections were made to item scoring for a small proportion of items. These corrections do not impact the fluency analyses. Fluency items were scored correctly and the overall scores did not change appreciably in mathematics.

Basic item analysis. The first question that we investigated was whether the placement and timing of the MF items made any appreciable difference in student performance on these

[7] The implication from the administrators' complaint being that, because the test was "too difficult", it was also too time-consuming for some students who spent an inordinate amount of time struggling with unfamiliar content and had little chance of reaching a correct response.

items. Tables 30 through 33 show the percent answering correctly (p-value) and the correlation of the item score with the total for all mathematics items in the form for Grades 3 through 6 respectively. Note that for Grades 3 and 6 all of the MF items were embedded within a single form, while for Grades 4 and 5 a smaller number of MF items were embedded within two different forms.

Table 30. Item P-Values and Item-Total Correlations for Grade 3 Math Fluency Items

Item	P-Values			Item-Total Correlations		
	Condition:			Condition:		
	End	End-Timed	Interspersed	End	End-Timed	Interspersed
1	0.26	0.28	0.20	0.54	0.55	0.52
2	0.46	0.45	0.42	0.60	0.62	0.60
3	0.72	0.73	0.73	0.40	0.37	0.38
4	0.47	0.44	0.46	0.61	0.60	0.59
5	0.76	0.80	0.75	0.43	0.42	0.41
6	0.66	0.63	0.67	0.52	0.49	0.49
7	0.48	0.50	0.54	0.57	0.57	0.58
8	0.13	0.15	0.11	0.48	0.53	0.49
Average	0.49	0.50	0.49	0.52	0.52	0.51

Table 31. Item P-Values and Item-Total Correlations for Grade 4 Math Fluency Items

Item	P-Values			Item-Total Correlations		
	Condition:			Condition:		
	End	End-Timed	Interspersed	End	End-Timed	Interspersed
A1	0.34	0.32	0.28	0.27	0.28	0.29
A2	0.54	0.53	0.56	0.46	0.44	0.41
A3	0.39	0.35	0.34	0.31	0.29	0.22
A4	0.25	0.25	0.25	0.39	0.37	0.37
A5	0.34	0.31	0.27	0.31	0.29	0.29
A6	0.36	0.34	0.31	0.30	0.29	0.29
B1	0.35	0.41	0.36	0.38	0.42	0.43
B2	0.49	0.52	0.56	0.44	0.50	0.41
B3	0.67	0.69	0.72	0.40	0.33	0.28
B4	0.37	0.30	0.31	0.38	0.29	0.29
B5	0.32	0.26	0.24	0.41	0.30	0.31
B6	0.37	0.32	0.30	0.28	0.18	0.22
Average	0.40	0.38	0.38	0.36	0.33	0.32

Table 32. Item P-Values and Item-Total Correlations for Grade 3 Math Fluency Items

Item	P-Values			Item-Total Correlations		
	Condition:			Condition:		
	End	End-Timed	Interspersed	End	End-Timed	Interspersed
A1	0.55	0.54	0.55	0.52	0.54	0.50
A2	0.25	0.23	0.20	0.47	0.45	0.44
A3	0.49	0.47	0.46	0.40	0.41	0.43
A4	0.18	0.19	0.21	0.46	0.51	0.46
A5	0.19	0.16	0.19	0.48	0.44	0.46
B1	0.20	0.20	0.22	0.36	0.44	0.44
B2	0.24	0.22	0.24	0.42	0.41	0.46
B3	0.16	0.17	0.16	0.33	0.42	0.38
B4	0.17	0.18	0.19	0.39	0.44	0.39
B5	0.26	0.25	0.27	0.45	0.46	0.48
Average	0.27	0.26	0.27	0.43	0.45	0.44

Table 33. Item P-Values and Item-Total Correlations for Grade 6 Math Fluency Items

Item	P-Values			Item-Total Correlations		
	Condition:			Condition:		
	End	End-Timed	Interspersed	End	End-Timed	Interspersed
1	0.18	0.22	0.21	0.38	0.32	0.42
2	0.06	0.09	0.07	0.29	0.23	0.32
3	0.53	0.52	0.50	0.51	0.50	0.50
4	0.39	0.44	0.41	0.28	0.32	0.39
5	0.22	0.23	0.29	0.42	0.37	0.45
6	0.57	0.56	0.61	0.43	0.38	0.35
7	0.26	0.27	0.27	0.40	0.43	0.42
Average	0.32	0.33	0.34	0.39	0.36	0.41

For Grade 3, the p-values for the first and last item were at or below chance levels. Three of the Grade 3 items were answered correctly by two-thirds of the students or more, which is more in line with what one might expect for a fluency item. However, only one of the 12 Grade 4 items and none of the Grade 5 or Grade 6 items met the two-thirds criterion. One of the Grade 6 items had a p-value less than .10 in all three conditions.

Item-total correlations for Grade 3 averaged around .5. Item total correlations for the other grades were a bit lower, but the correlations were above .2 for all items (averaged across condition). The items each appeared to be a consistent part of the overall mathematics measure, although these correlations may have been inflated somewhat by a motivation factor. Students who tried hard on some items were likely to try harder on most or all of the items, inflating the inter-item correlations.

Tables 30 - 33 indicate that the placement, timing, and instructions did not make any noticeable differences in the basic item statistics. Average p-values and item-total correlations were nearly identical across the three testing conditions for each grade.

Given the varying difficulty and reasonable item-total correlations for the MF items, we computed MF number correct scores. Figures 4 -7 show the distribution (frequencies) of these number correct scores for Grades 3 through 6 respectively.

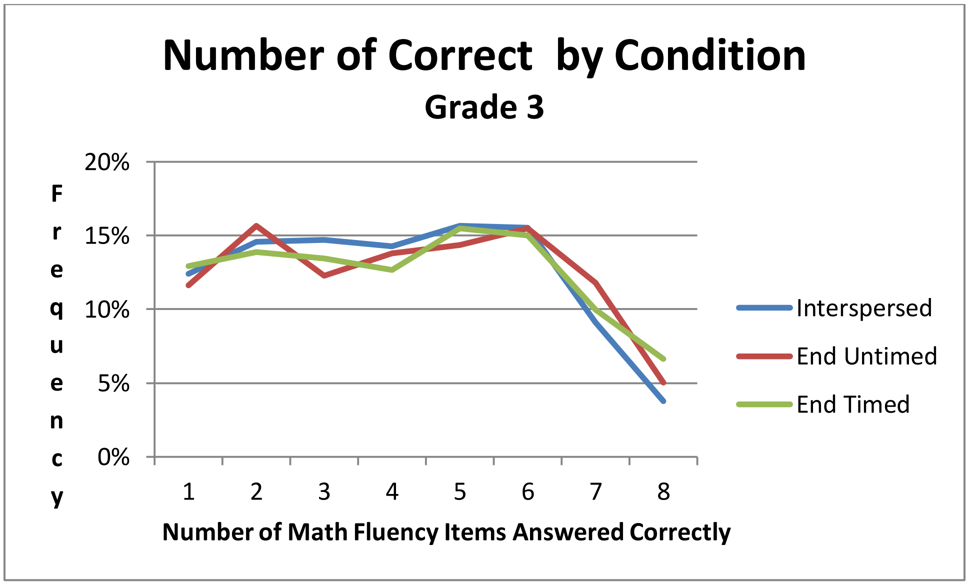


Figure 4. Percentage of grade 3 students at each number correct score level (excluding students with no correct responses).

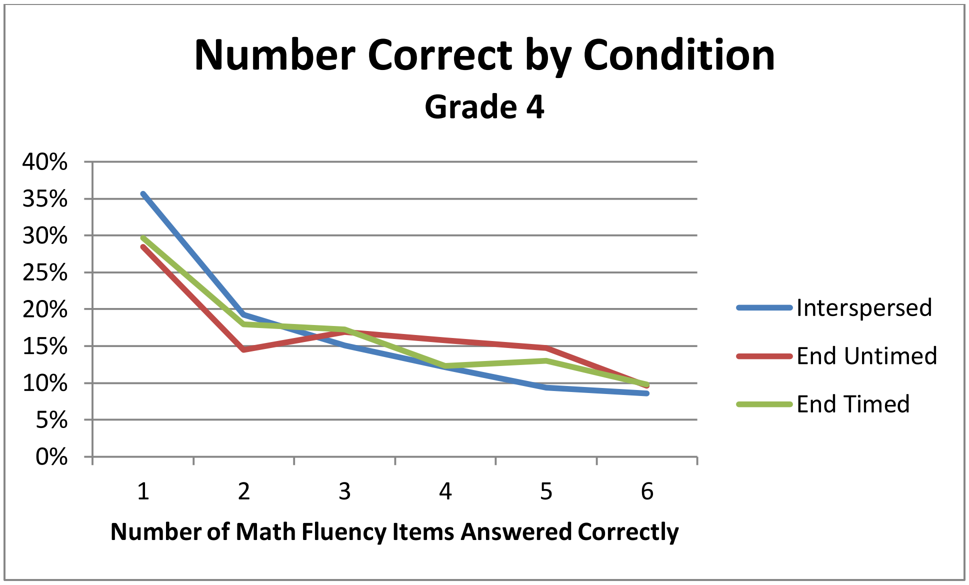


Figure 5. Percentage of grade 4 students at each number correct score level (excluding students with no correct responses).

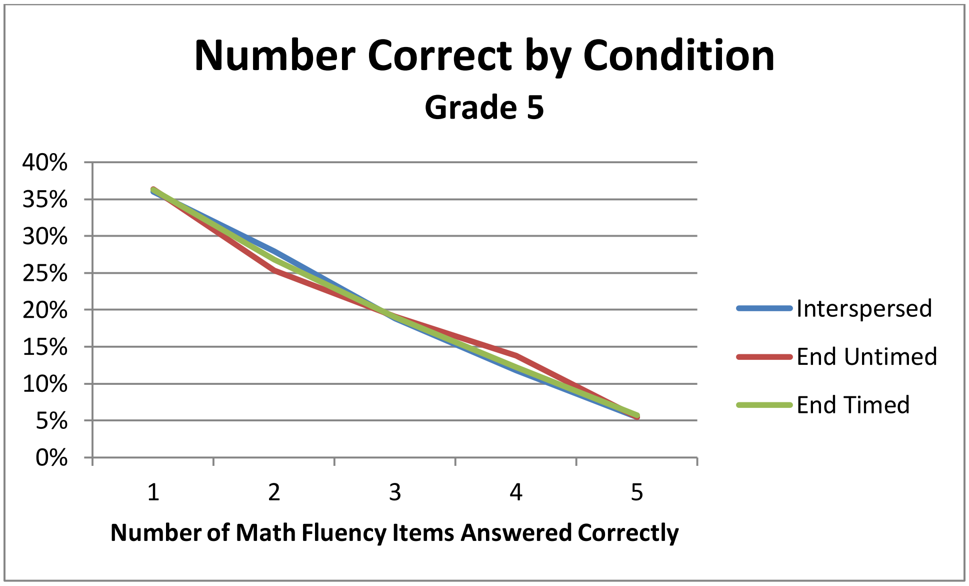


Figure 6. Percentage of grade 5 students at each number correct score level (excluding students with no correct responses).

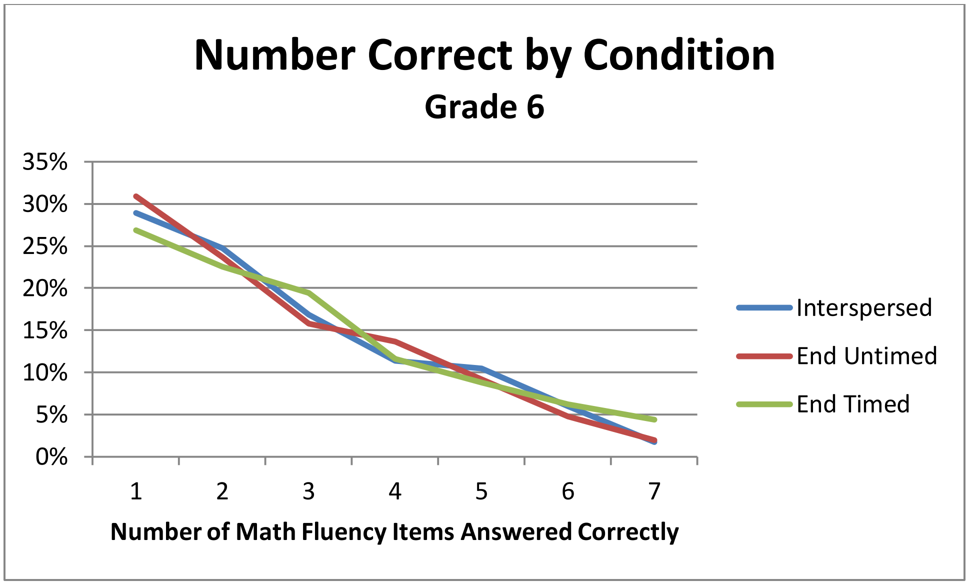


Figure 7. Percentage of grade 6 students at each number correct score level (excluding students with no correct responses).

As indicated by Figures 4 - 7, placement, timing, and instructions did not make any noticeable differences in the total number of MF items answered correctly.

Analysis of timing data. We used timing data to construct an initial indicator of mathematics fluency (MF), defined as the average of the log of the response times for correct responses to the math fluency items. Obviously, this measure was undefined if there were no correct responses, and the measure may not be as meaningful as the number correct score for students who do not answer many or most of the items correctly.

Figures 8 - 11 show the average timing-based fluency indicator by number correct and testing condition (MF items interspersed, at the end, and at the end with separate timing instructions). For students answering fewer than half of the MF items correctly, those who answered faster tended to have lower number correct scores. It is likely that these students were either not trying or did not have the requisite skill and were just guessing quickly. For students who answered half or more of the MF items correctly, students with higher number correct scores tended to answer slightly more rapidly.

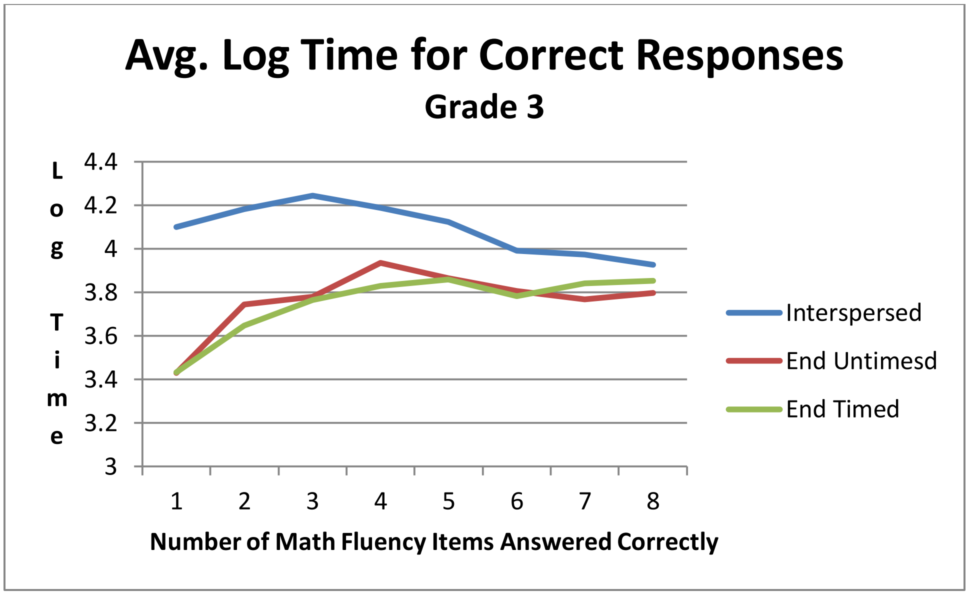


Figure 8: Math Fluency indicator by condition and number correct for Grade 3 students.

The number correct distribution is, perhaps surprisingly, very similar when the items are interspersed as when they are at the end. Also, separate timing and instructions do not appear to affect the number correct distribution.

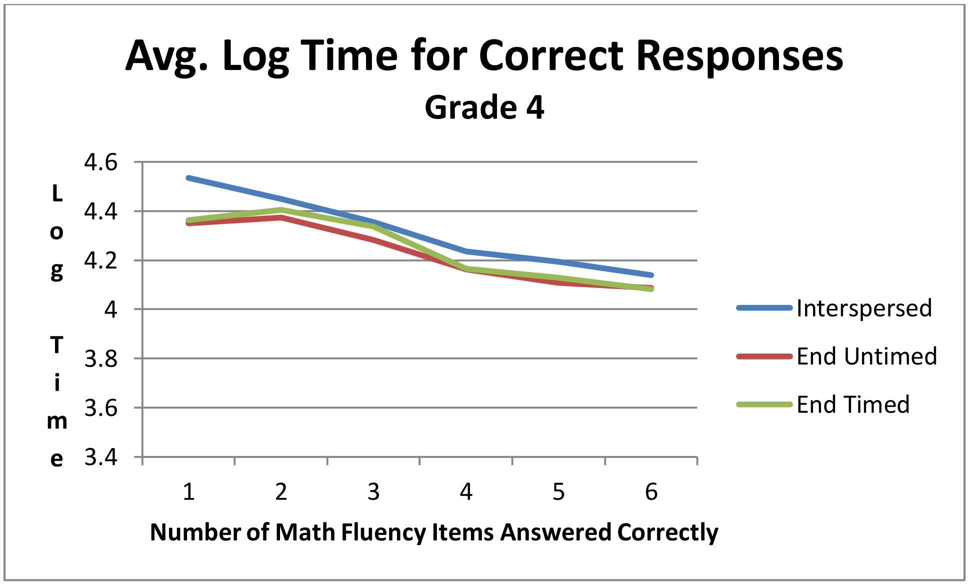


Figure 9: Math Fluency indicator by condition and cumber correct for Grade 4 students.

Grade 4 students tend to spend a little more time per item in comparison to grade 3 students and there is a consistent trend of decreasing average time with increasing number correct. Although there is not much difference by condition in the number correct distributions, relatively fewer Grade 4 students are answering most of the MF items correctly compared to Grade 3. Further analyses may be needed to see if the Grade 4 items are harder or flawed or if the Grade 4 students are just less motivated.

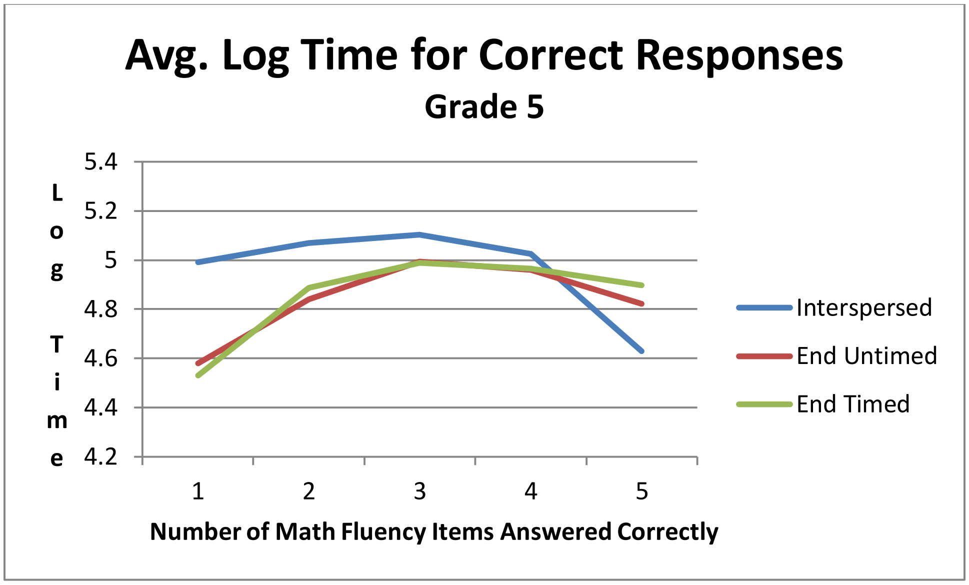


Figure 10: Math Fluency indicator by condition and cumber correct for Grade 5 students.

Grade 5 students spend yet a bit more time per item. The relationship of time spent to number correct is more unimodal than for either Grade 3 or Grade 4. Again, average time decreases with number correct for students answering more than half of the items correctly. Grade 5 students also had lower number correct distributions than even the Grade 4 students, again indicating some issue with the items or with the motivation of the students taking the items.

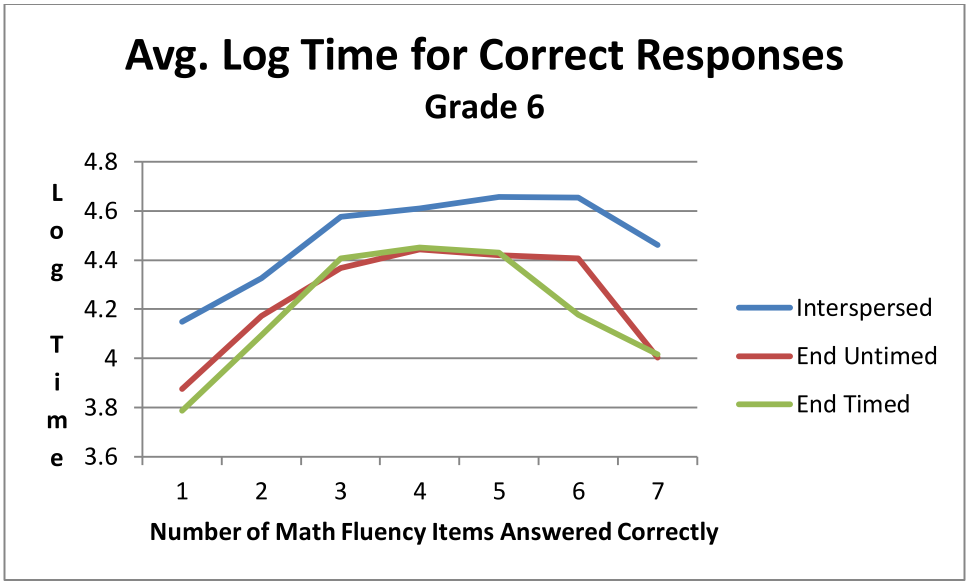


Figure 11: Math Fluency indicator by condition and cumber correct for Grade 6 students.

The relationship between time per item and number correct is also unimodal for Grade 6 students, although students are spending somewhat less time per item compared to Grade 5 students, particularly those not getting many right. Similar to the other grades, Grade 6 students did not do very well on the math fluency items.

Reliability. Options for using the field test data to estimate the reliability of various math fluency scores are limited. First, true score differences in fluency are confounded with examinee differences in test taking motivation. We used a split-half approach to provide preliminary reliability estimates. A split-half approach was needed rather than an item-by-item internal consistency estimator (Coefficient Alpha) because the response time indicator was only defined for items that were answered correctly. Thus, a given student would have a response time indicator for some items but not for others. Most students would, however, have response time data for at least one of a group (half form) of items. Examining the correlation of the response time indicators across two separate sets of items was thus selected to provide a first approximation to a reliability estimate. While these estimates are no doubt inflated, they do provide a basis for comparing different indicators, as will be shown.

Appendix G shows detailed results from the analyses of split-half data. For each grade, condition and form, the tables show the number of students with usable number correct (NC) and log correct response time (LCT) scores, the NC and LCT means for the odd and even halves, the correlations of the odd and even half scores, and the estimated reliability for the full-length test. The halves were short, two to four items each. In the first set of analyses, all students were included (with no further screening) for the number correct (NC) analyses and all students with at least one correct response to a math fluency item were included in the response time analyses. As shown, this resulted in somewhat smaller sample sizes for the response time analyses. Across all grades and conditions, the reliability estimate for the full-length (5- to 8-item) math fluency test was .93. As noted, this is an incredibly high estimate for such a short test, but the data are the data. The average estimated reliability for the response time measure was .53.

The second table in Appendix G shows results for the same analyses using only students who answered half or more of the math fluency items correctly (4 out of 7 or 8, for grades 3 and 6; and 3 out of 5 or 6, for grades 4 and 5). As expected, given the restricted range of the NC scores, the reliability estimate was lower but still surprisingly high, averaging .86. While the reliability estimate for the number correct scores decreased, the reliability estimate for the response time scores actually increased to .81. Within this range of students who clearly had some skill with math fluency problems, the response time indicator was nearly as consistent across items as was the dichotomous item scores.

Table 34 shows the NC and LCT reliability estimates by testing condition for all students and for students answering at least half of the MF items correctly (Selected). The NC reliability estimates were virtually identical across the testing conditions. The LCT reliability estimates were slightly higher for the separately timed condition and somewhat lower when the items were interspersed throughout the test, but these differences were small, even for the selected student sample.

Table 34. Reliability Estimates by Testing Condition for All and for Selected Students

	All Students		Selected Students	
Condition	Number Correct	Log Correct Response Time	Number Correct	Log Correct Response Time
End of Test	.93	.54	.86	.81
End and Timed	.93	.54	.87	.83
Interspersed	.92	.51	.86	.78
Average	.93	.53	.86	.81

Correlation with other measures. Finally, we examined the relationship between proficiency measures using the MF items and measures of proficiency in other areas of mathematics. We created an "Other Math" score by subtracting the MF number correct score from the total number correct score for each of the field test forms containing MF items.

Tables 35 and 36 show results of regression analyses using both the MF NC and LCT scores to predict each student's score on the other mathematics items, for all students and for higher ability (at least half of the MF items answered correctly) respectively.

Table 35. Regression Analysis Results for All Students

Stat	Var	Grades			
		3	4	5	6
No. of MF Items		8	6	5	7
N		3947	5951	4528	2643
Means	NC	4.16	2.84	2.24	2.78
	LCT	3.92	4.32	4.89	4.24
	Other	22.32	21.54	21.15	18.29
SDs	NC	2.07	1.69	1.22	1.67
	LCT	0.64	0.44	0.82	0.84
	Other	10.23	10.87	9.77	9.58
Corr. With Other	NC	0.74	0.25	0.41	0.44
	LCT	-0.04	-0.20	0.15	0.23
Reg. Coef. T-Values	NC	68.96	16.34	30.03	23.57
	LCT	-5.73	-10.67	8.67	8.43
R-Square		0.55	0.08	0.18	0.22
Increase in R-Square		0.00	0.02	0.01	0.02

Table 36. Regression Analysis Results for High Scoring Students

Stat	Var	Grades			
		3	4	5	6
No. of MF Items		8	6	5	7
N		2341	3014	1672	797
Means	NC	5.62	4.28	3.64	4.97
	LCT	3.94	4.21	4.98	4.45
	Other	27.57	23.84	25.92	24.19
SDs	NC	1.24	1.10	0.73	0.98
	LCT	0.49	0.36	0.46	0.49
	Other	8.74	10.86	9.47	10.11
Corr. With Other	NC	0.57	0.20	0.14	-0.01
	LCT	-0.20	-0.18	-0.01	0.29
Reg. Coef. T-Values	NC	32.74	8.92	5.99	NS
	LCT	-7.76	-7.70	N.S.	8.55
R-Square		0.34	0.06	0.02	0.08
Increase in R-Square		0.02	0.02	0.00	0.08

Grade 3 students appeared to take the MF items seriously (or had more skill in the tested mathematics). More than half of the Grade 3 students answered 50 percent or more of the MF correctly (2,341 in Table 36 compared to 3,947 in Table 35; also the unrestricted mean of 4.16 for the Grade 3 NC scores in Table 35 is above 50 percent). This was not the case for Grades 4 through 6 where the unrestricted mean NC scores were well below 50 percent of the total number of items. Also, Grade 3 NC scores were highly correlated with the Other Math score (.74), while NC scores for Grades 4 through 6 were not (.25 to .44).

A second observation is that the LCT scores were negatively correlated with the other math score for Grades 3 and 4 as expected – quicker response times indicating greater ability. For the restricted sample, these correlations were about -.2. For Grades 5 and 6, however, the LCT scores had a positive correlation with the other math score. In this case, spending more time on each item may have been an indicator of greater effort.

Finally, the regression results for Grades 3 and 4 did indicate that using both NC and LCT scores to predict the other math score improved prediction, albeit very modestly, compared to using NC scores alone.

Findings for Research Question 4: Do rubrics lend themselves to accurate and reliable scoring?

Two sets of analyses were conducted to address research question 4. First, observational data from the scoring site visits was analyzed to determine if the training and qualification requirements and the operational scoring procedures supported the accurate and reliable scoring of hand-scored item responses. Next, a set of alternate scoring rules was applied to a sample of partial credit items to investigate how different item components contribute to the overall item score.

Scoring site visits. Two HumRRO staff observed the scorer training in June 2014. During this scorer training visit, HumRRO staff separately observed two of three training groups, each comprised of more than 35 experienced scorers. On a second visit in June-July 2014, the same HumRRO staff members separately observed live scoring by two pods. Following these observations, HumRRO staff discussed their notes and reached consensus on their findings and recommendations.

In general, for the tasks HumRRO researchers observed, scoring procedures were found to lend themselves to accurate and reliable scoring insofar as they were consistent with the expectations and activities described in the PARCC Handscoring Specifications 4.1 document. It is important to note that the sample was very small for both visits and HumRRO observers were only available to observe 2 days each visit, so not all processes were observed. Therefore, findings are not necessarily generalizable to the full range of scoring activities.

During the scorer training visit, it was noted that the centerpiece of training is the scoring rubric. Scoring trainers were observed methodically discussing all score points for each applicable construct of the rubric. A thorough review of anchor papers was also observed, and student responses in the anchor paper set were annotated with the rationale of the assigned score. After completing the anchor paper review, Scoring Directors handed out the first set of practice papers (approximately 14) and emphasized using the rubric and anchor papers to make scoring decisions. As a calibration activity, just the first 3 or 4 papers in the practice set were scored individually and then the rationale for the correct scores was discussed. The remaining papers in the set were then scored and reviewed. Due to the length of this practice period, HumRRO observers were not able to view the completion of the practice paper task and the administration of the qualifying sets.

During the item scoring site visit, HumRRO researchers observed scoring for two writing assessment items. In general, observers found scoring to be consistent with procedures established in the Handscoring Specifications 4.1 document. Throughout item scoring, the Scoring Director and Supervisor monitored scoring statistics and quality checks as required. Some scorers did have difficulty applying the rubric consistently across student responses and required individual calibration training. The Scoring Directors or Supervisors worked with those scorers individually to diagnose the source of their difficulty. This was accomplished by working together to identify similar student responses in the anchor and practice papers and then reviewing the annotations of the scoring rationales. Several times, scorers were unable to make the adjustments needed in their scoring decisions and, after required warnings and recalibrations, were released. Scoring Directors and Supervisors were observed accessing, reviewing, and using ePEN reports as prescribed in the handscoring specifications. It was noted by both observers that multiple steps are required to access various scoring reports. Although the steps were not difficult to complete, due to the number of times reports were checked, it became apparent that it could become frustrating over time. This was particularly evident for ribbed items. Ribbed items were included in both the paper and pencil assessment as well as the computer-based assessment. It was observed that Scoring Directors had to access productivity and item quality reports separately and that scoring statistics for paper and computer-based were not tracked together. The only process not observed during the site visit was item closeout; however, one HumRRO observer confirmed item documentation and archive procedures with a Scoring Director.

Alternate scoring rules. The PARCC assessments included several two-part items that were scored using a set of scoring rules. Following the PARCC Rules Based Scoring Meeting in Iowa June30-July1, 2014, PARCC requested that certain items be rescored using a set of alternate scoring rules. This included two-part technology enhanced items and evidence-based selected response items with three or more Evidences, as well as two-part technology enhanced items with six or more student responses in the Accuracy portion. Table 37 presents the original and alternate scoring rules for the two item types.

Table 37. Original and Alternate Scoring Rules for Two-Part Items

Original Rules			Alternate Rules		
2-part technology enhanced items & evidence-based selected response items with 3+ Evidences					
Accuracy	**Evidence**	**Score**	**Accuracy**	**Evidence**	**Score**
Fully correct	Fully correct	2	Fully correct	Fully correct	2
Fully correct	Partially correct	1	Fully correct	Partially correct (N-1)	2
Fully correct	Incorrect	1	Fully correct	Partially correct (<N-1)	1
Partially correct	Fully correct	1	Partially correct	Fully correct	1
Partially correct	Partially correct	1	Partially correct	Partially correct	1
Partially correct	Incorrect	1	Partially correct	Incorrect	1
Incorrect	Fully correct	0	Incorrect	Fully correct	0
Incorrect	Partially correct	0	Incorrect	Partially correct	0
Incorrect	Incorrect	0	Incorrect	Incorrect	0
2-part technology enhanced items with 6+ student responses in the Accuracy portion					
Accuracy	**Evidence**	**Score**	**Accuracy**	**Evidence**	**Score**
Fully correct	Fully correct	2	Fully correct	Fully correct	2
Fully correct	Partially correct	1	Fully correct	Partially correct	1
Fully correct	Incorrect	1	Fully correct	Incorrect	1
Partially correct (n-1)	Fully correct	1	Partially correct (N-1)	Fully correct	2
Partially correct (n-1)	Partially correct	1	Partially correct (<N-1)	Full correct	1
Partially correct (n-1)	Incorrect	1	Partially correct	Incorrect	1
Incorrect	Fully correct	0	Incorrect	Fully correct	0
Incorrect	Partially correct	0	Incorrect	Partially correct	0
Incorrect	Incorrect	0	Incorrect	Incorrect	0

As Table 37 shows, the two item parts reflected Accuracy and Evidence components, respectively. Depending on the number of response options that a student was prompted to select in each item part, a given response could be deemed 'fully correct,' 'partially correct,' or 'incorrect.' Points were awarded based on the combinations of these judgments. The upper portion of Table 37 presents scoring rules for items in which students were prompted to provide three or more pieces of evidence in their response (i.e., the number of pieces of evidence that students were prompted to provide was item-specific, but was at least three). The lower portion presents scoring rules for items in which students were prompted to provide six or more responses in the Accuracy component of the item.

Comparing the original and alternative rules for both item types shows that the alternative rules provided additional opportunities to obtain scores of 2 and 1. For example, under the alternative rules for items with three or more evidences, students could receive a score of 2 if they provided all but one of the correct pieces of evidence, which would have rendered a score of 1 under the original scoring rules.

It is important to note that HumRRO staff went through a series of steps to check the accuracy of the rescoring methods applied. First, analysts worked to match the original item scores in the data file to make sure item parts were being scored correctly. Three items with potential scoring issues were identified and reviewed and the sources of the errors were identified for two of the three items. We proceeded by using the HumRRO-generated score as the original score for these three items. A few additional cases were identified where students provided more responses than an item prompted them for. These cases were removed from additional part analyses since there were some concerns on how to score these items. After HumRRO staff was satisfied with the match of the original scores based on part item scores, the new scoring rules were applied.

Figures 12 and 13 provide examples of the effects of the alternative scoring rules on the distributions of scores. The first figure depicts the distribution of scores for an item in which three or more evidences were to be included in the student response. Figure 12 reflects that the alternate scoring rules provided an additional opportunity for scoring a 2 and shows that the number of 0 scores did not change, while the number of 1 scores decreased and number of 2 scores increased with the application of alternate scoring rules. The second figure depicts the distribution of scores for an item in which six or more responses to the Accuracy component were to be included in the student response. Here, the alternate scoring rules provided more opportunities for scores of both 1 and 2, as reflected in the decrease in 0 scores and increases in 1 and 2 scores. Similar graphs for each item are presented in Appendix H.

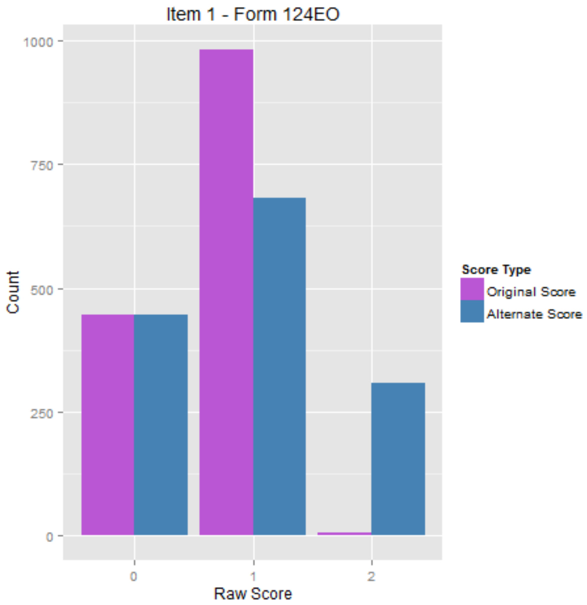


Figure 12. Effects of alternate scoring rules on score distribution of item with three or more evidences.

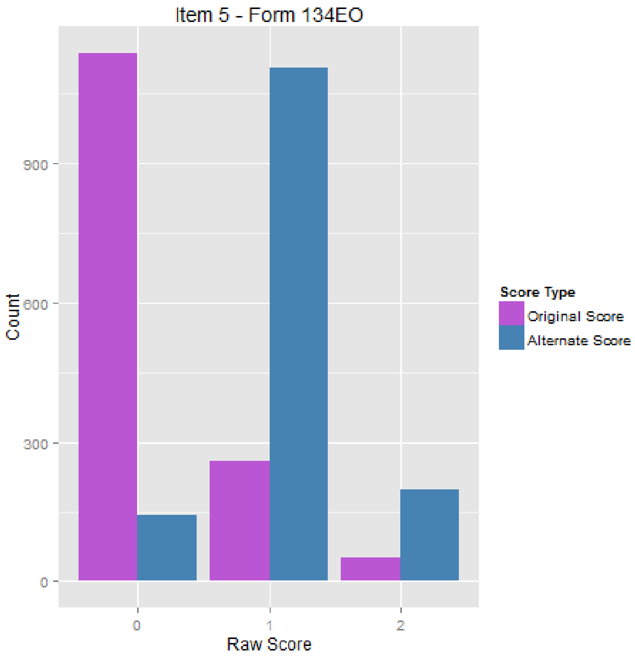


Figure 13. Effects of alternate scoring rules on score distribution of item with six or more responses in the Accuracy component.

Table 38 presents point-biserial item-total correlations for a sample of two-part items scored using both sets of rules. Item-total correlations are one way to demonstrate how well items are discriminating between high-and low-performing students. For example, a low item-total correlation indicates that high-and low-performing students are similar in their probabilities of answering an item correctly. The leftmost column presents the form and booklet position of each item included in this analysis. The middle column presents the correlation between the item score, using the original scoring rules, and the total score (with the contribution of the item removed). The rightmost column presents the correlation between the item score and the total score (with the contribution of the item removed) when the alternative scoring rules were applied.

Table 38 shows that in general, changes in the item-total correlations were minimal after applying the alternative scoring rules, indicating that this alternative approach to scoring two-part items did not substantially improve the items' abilities to distinguish between high- and low-performing students. There was one item (Grade 8, Form024EO, Item 32) for which the original scoring rules yielded a very small, negative item-total correlation, but for which the alternative scoring rules yielded a small, positive correlation. A closer look at this item indicated that under the original scoring rules, students were required to select at least five out of six correct responses to receive at least a score of 'partially correct', something which the large majority of students failed to do.

Table 38. Item-Total Correlations Under Two Sets of Scoring Rules

	Original	Alternative
Grade 5		
014EO Item 5	.31	.29
124EO Item 1	.36	.32
124EO Item 25	.25	.23
054EP Item 27	.02	.09
134EO Item 4	.17	.16
134EO Item 5	.35	.33
Grade 8		
024EO Item 2	.21	.17
024EO Item 4	.14	.14
024EO Item 12	.22	.24
024EO Item 18	.45	.46
024EO Item 32	-.08	.22
024EP Item 4	.50	.50
024EP Item 6	.30	.27
024EP Item 12	.48	.52
064EO Item 8	.22	.25
064EO Item 16	-.20	-.20
064EO Item 17	.44	.52
064EO Item 25	.49	.47
064EO Item 36	.24	.29
094EO Item 4	.09	.13
094EO Item 6	.38	.41
094EO Item 8	.22	.27
094EO Item 20	.27	.30
094EO Item 25	.27	.30
094EO Item 29	.40	.50

Another way to look at the information provided by multi-part items is to analyze the correlation between each component of the item and the total test score. This provides some indication of which component of the item is most useful for differentiating between students of differing levels of overall performance. Table 39 indicates that in general, the Evidence components tend to have somewhat higher item-total correlations. This is not surprising given that the Evidence subscore used in calculating these correlations tended to have a larger range of possible scores (i.e., three or more points). There were two items for which the Accuracy subscore had an essentially zero or a negative item-total correlation. The negative item-total correlation is particularly concerning as it would indicate that students who were higher scoring overall were slightly more likely to answer the Accuracy component of this item (Grade 8, Form 064EO, Item 16) incorrectly.

Table 39. Item-Total Correlations of Rescored Item Parts

	Accuracy	Evidence
Grade 5		
3+ Evidences		
014EO Item 5	.29	.20
124EO Item 1	.36	.12
124EO Item 25	.20	.35
054EP Item 27	.001	.28
6+ Accuracies		
134EO Item 4	.16	.36
134EO Item 5	.22	.37
Grade 8		
3+ Evidences		
024EO Item 2	.12	.31
024EO Item 4	.14	.19
024EO Item 12	.18	.20
024EO Item 18	.40	.48
024EP Item 4	.48	.46
024EP Item 6	.19	.25
024EP Item 12	.45	.44
064EO Item 8	.17	.39
064EO Item 16	-.20	.21
064EO Item 17	.44	.40
064EO Item 25	.43	.46
064EO Item 36	.25	.27
094EO Item 4	.07	.34
094EO Item 6	.41	.30
094EO Item 8	.18	.44
094EO Item 20	.28	.15
094EO Item 25	.26	.44
094EO Item 29	.46	.41
6+ Accuracies		
024EO Item 32	.28	.34

Table 40 further explores the relations between item parts by presenting the correlations between the accuracy and evidence components for each item. Although the correlations are reasonably high for several items, over 60% of the items included had correlations between item parts that were lower than .30. Inspection of the item characteristic curves produced when item parts were treated separately (see Appendix I) indicates that for some items, examinees did not receive points for Accuracy, but did receive some points for Evidences. Patterns such as this would explain these low correlations, and would raise additional concerns about loss of information caused by combining the parts into a single item.

Table 40. Correlations between Item Parts

Form and Item	Correlation
Grade 5	
014EO Item 5	0.17
124EO Item1	0.10
124EO Item25	0.10
134EO Item 4	0.14
134EO Item 5	0.20
134EP Item 27	0.18
Grade 8	
024EO Item 12	0.28
024EO Item 18	0.54
024EO Item 2	0.25
024EO Item 32	0.21
024EO Item 4	0.53
024EP Item 12	0.32
024EP Item 4	0.64
024EP Item 6	0.68
064EO Item 16	0.00
064EO Item 17	0.36
064EO Item 25	0.50
064EO Item 36	0.19
064EO Item 8; 094EO Item 8	0.22
094EO Item 20	0.19
094EO Item 25	0.15
094EO Item 29	0.47
094EO Item 4	0.05
094EO Item 6	0.37
024EO Item 12	0.28
024EO Item 18	0.54
024EO Item 2	0.25
024EO Item 32	0.21
024EO Item 4	0.53
024EP Item 12	0.32

A final comparison of item parameters required Item Response Theory (IRT) estimation to complete. Using the full set of items on the two indicated forms, the two-parameter-partial credit (2PPC) model was used to estimate slope (discrimination) and cross-over points (the point on the theta scale where the likelihood of receiving a lower score become less likely than the next higher score). Item slopes and crossover points provide additional information regarding the potential impact of alternate scoring rules, or of treating item parts separately. Table 41 summarizes this information for all included items (crossover points on the theta scale), while Figure 14 provides an example visual representation of the items (similar figures for all included items are presented in Appendix I). The table and figures demonstrate that although the application of alternate scoring rules tends to result in a shift to the left toward a more normal distribution, the non-ascending crossover point values still suggest that items may not be functioning in the intended ways. Treating the item parts separately tends to yield crossover point values that follow the intended pattern.

Table 41. Rescored Item Information

			Crossover Points						
Form/Item	**Score**	**Slope**	**0/1**	**1/2**	**2/3**	**3/4**	**4/5**	**5/6**	**6/7**
Grade 5									
014EO Item 5	original	0.25	0.00	1.73	0.20				
	alternate	0.20	0.00	5.59	-4.87				
	part 1	0.38							
	part 2	0.18	0.00	-7.22	-2.69	-0.36			
124EO Item 1	original	0.60	0.00	-0.96	5.55				
	alternate	0.32	0.00	-0.98	1.70				
	part 1	0.60							
	part 2	0.11	0.00	-2.13	2.01	21.91			
124EO Item 25	original	0.26	0.00	1.37	3.12				
	alternate	0.17	0.00	7.69	-5.81				
	part 1	0.27							
	part 2	0.40	0.00	-2.73	-2.49	1.85			
134EO Item 4[a]	original	0.46							
	alternate	0.62							
	part 1	0.24	0.00	-8.50	1.86	-3.16	3.21	2.74	5.89
	part 2	0.25							
134EO Item 5	original	0.53	0.00	1.85	2.69				
	alternate	0.51	0.00	-2.65	2.32				
	part 1	0.18	0.00	-1.30	-0.98	-0.22	-1.06	1.53	4.68
	part 2	0.33							
054EP Item 27	original	0.08	0.00	5.58	24.62				
	alternate	0.09	0.00	8.58	1.55				
	part 1	0.08							
	part 2	0.30	0.00	-2.41	0.93	6.60			

(continued)

Table 41. Rescored Item Information (continued)

			Crossover Points						
Form/Item	Score	Slope	0/1	1/2	2/3	3/4	4/5	5/6	6/7
				Grade 8					
024EO Item 2	original	0.20	0.00	3.03	1.80				
	alternate	0.13	0.00	9.16	-5.48				
	part 1	0.19							
	part 2	0.30	0.00	-4.00	-1.62	1.92			
024EO Item 4[a]	original	0.24							
	alternate	0.24							
	part 1	0.26							
	part 2	0.35							
024EO Item 12	original	0.29	0.00	3.01	3.58				
	alternate	0.25	0.00	5.06	-0.40				
	part 1	0.31							
	part 2	0.18	0.00	-3.25	1.76	4.77			
024EO Item 18	original	0.57	0.00	0.18	1.82				
	alternate	0.46	0.00	1.40	-0.85				
	part 1	0.73							
	part 2	0.52	0.00	-1.12	0.00	1.72			
024EO Item 32	original	0.06	0.00	28.68	19.74				
	alternate	0.62	0.00	-3.30	3.80				
	part 1	0.27	0.00	2.33	-4.93	-0.78	-5.10	6.21	3.05
	part 2	0.41							
024EP Item 4	original	0.85	0.00	-1.76	1.42				
	alternate	0.73	0.00	-1.82	0.78				
	part 1	0.80	0.00	-1.72	0.59				
	part 2	0.51	0.00	-1.90	0.32	0.85			
024EP Item 6	original	0.25	0.00	2.69	-0.14				
	alternate	0.21	0.00	4.75	-2.57				
	part 1	0.31							
	part 2	0.18	0.00	0.60	2.57	-1.62			
024EP Item 12	original	0.63	0.00	0.13	1.73				
	alternate	0.55	0.00	0.83	-0.17				
	part 1	0.80							
	part 2	0.45	0.00	-1.59	0.07	1.50			
064EO Item 8; 094EO Item 8	original	0.22	0.00	-0.83	2.67				
	alternate	0.20	0.00	1.46	-2.19				
	part 1	0.24							
	part 2	0.26	0.00	-0.51	-3.14	-2.01	-1.51	1.41	

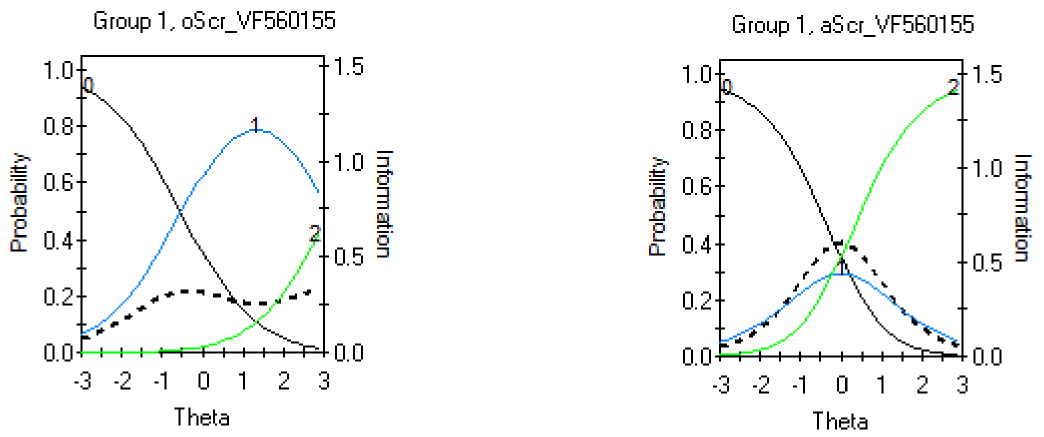
(continued)

Table 41. Rescored Item Information (continued)

			Crossover Points						
Form/Item	**Score**	**Slope**	**0/1**	**1/2**	**2/3**	**3/4**	**4/5**	**5/6**	**6/7**
Grade 8 (continued)									
064EO Item 16[a]	original	0.04							
	alternate	0.03	0.00	47.76	34.30				
	part 1	0.04							
	part 2	0.22	0.00	-2.38	3.38				
064EO Item 17	original	0.64	0.00	-0.52	3.15				
	alternate	0.54	0.00	0.20	-0.21				
	part 1	0.79							
	part 2	0.45	0.00	-2.21	-0.11	3.58			
064EO Item 25	original	0.48	0.00	1.37	-0.02				
	alternate	0.40	0.00	3.17	-2.33				
	part 1	0.74							
	part 2	0.46	0.00	-2.45	-0.60	0.38			
064EO Item 36	original	0.26	0.00	0.76	10.43				
	alternate	0.19	0.00	4.18	-1.78				
	part 1	0.29							
	part 2	0.30	0.00	-3.43	-0.75	7.11			
094EO Item 4	original	0.13	0.00	4.52	13.40				
	alternate	0.15	0.00	4.73	5.84				
	part 1	0.11							
	part 2	0.23	0.00	-1.52	1.49	1.86	3.59		
094EO Item 6	original	0.43	0.00	0.04	2.11				
	alternate	0.34	0.00	2.73	-2.73				
	part 1	0.64							
	part 2	0.31	0.00	-3.15	-1.80	2.55			
094EO Item 20	original	0.37	0.00	1.55	4.30				
	alternate	0.30	0.00	2.86	0.81				
	part 1	0.43							
	part 2	0.15	0.00	-4.35	1.89	7.11			
094EO Item 25	original	0.30	0.00	0.73	4.44				
	alternate	0.24	0.00	2.87	-0.87				
	part 1	0.33							
	part 2	0.44	0.00	-2.06	-0.43	2.76			
094EO Item 29	original	0.44	0.00	-0.98	1.70				
	alternate	0.49	0.00	1.00	-2.30				
	part 1	0.85							
	part 2	0.43	0.00	-1.83	-1.84	1.62			

[a] The response categories for these items were collapsed during psychometric analyses due to insufficient numbers of responses at a particular score point. The particular score points which were collapsed are as follows: 134EO Item 4 original point 2, alternate point 0, and Part 1 point 0 and 1;   024EO Item 4 original and alternate point 2, and Part 2 points 2,3,and 4; and 064EO Item 16 original point 2 and Part 2 point 3.

Figure 14 presents a graphical representation of the same item under the original and alternate scoring rule. The dark solid line represents the probability of receiving 0 points on the item by theta. The blue line represents the probability of receiving 1 point and the green line the probability of receiving 2 points. The dotted line indicates the item information curve (ICC). As can be seen in the graph, under the original scoring rule, the probability of getting 2 points does not occur until theta is more than three standard deviations beyond the mean. However, under the alternate scoring rule, there is no theta for which a score of 1 is most likely (0 and 2 lines cross instead).


Original Rule

Alternate Rule

Figure 14. Comparison of Item Function under the Original and Alternate Scoring Rules: Grade 8 Form 064EO, Item 17.

Summary and Recommendations

The overall summary and recommendations, based on the findings for the research questions, are discussed in this section by each claim in the TOA that was addressed.

Claim 1: The Assessment Connects with Common Core State Standards (CCSS)

To investigate the connection with the CCSS, we began by examining the distribution of items that were dropped for various reasons from analysis. It is not uncommon to lose a substantial number of items based on their performance during field test. This is why it is common to field test 25% or more additional items, beyond those needed for operational test construction. This practice allows testing companies to meet test content representation requirements even if some items do not survive field test. Problems can arise when a very large number of items fail to meet field test performance expectations, or if the items that fail to meet performance requirements are concentrated in a particular grade, subject, or sub-content area.

This study is based on field test items that were not dropped due to administrative and statistical concerns. Items were dropped for administrative reasons (identified by Pearson) and for statistical reasons (described in the body of this report). PARCC EOY assessments had between 13% and 42% of items dropped for these reasons. If we use 25% as a guide[8] for further scrutinizing an assessment, grade 3, 8, and 9 ELA, grade 6 Math, Algebra 1, Algebra 2, Geometry, and Integrated Math 1, 2, and 3 all had 25% or more items dropped from analyses. High school mathematics assessments were especially troubling because the proportion of dropped items was often greater than 30%, and greater than 40% for Integrated Math 2. . The PBA assessments had between 15% and 46% of items dropped. Only grade 3 among the ELA assessments topped 25% drops, but several math assessments had greater than 25% drops. Like the EOY assessments, both Integrated Math 2 and 3 had the largest number of dropped items, at 42% and 46% respectively. In addition, Math grade 5, 6, and 8, Algebra 1, Algebra 2,Geometry and Integrated Math 1 all had more than 25% dropped items. We expect that it may be difficult to support the planned number of test forms and maintain content coverage requirements for high school level mathematics without adding items to the current item pool. Current item development efforts should make high school mathematics, especially for the PBA, a priority.

Given the large proportion of potential items dropped for some tests, we recommend the analyses presented in this report be repeated with the full set of items on the operational exams. Further, while we expect that most of the items dropped in this analysis were not

[8] We would typically expect fewer than 25% of items to be dropped due to poor performance, but used this proportion as our acceptability criterion based on typical field testing practice and the minimum number of required operational items.

included in the operational test forms, we recommend any flagged items that were included in the 2015 tests be subject to further scrutiny. We did not track items from the field test to operational tests. If the items continue to perform poorly, we would recommend that PARCC consider removing them before generating operational student scores. Poor performing items, especially those with poor discrimination, can negatively impact test reliability as well as the validity of inferences from test scores. If items must be omitted, PARCC should be cautious regarding the adequacy of the test item pool to represent the intended construct, especially when reporting sub-scores.

Correlations with state scores provide considerable validity evidence for the PARCC PBA and EOY assessments in both ELA and math. The pattern of convergent/discriminant validity coefficients was as expected (convergent validity coefficients were higher than discriminant coefficients, but all were positive) for the large majority of correlations across five states. Correlations were strong for both reading and math for grades 3-8, but less so for high school. This provides promising evidence that the construct measured by PARCC is not greatly different from that measured by the state assessments, at least for the elementary and middle grades, but the correlations were not so strong as to indicate that the tests are redundant. There are several potential reasons why the high school correlations were poorer than lower grade correlations. First, high school students tend to display more motivation effect than younger students and there were no stakes for scoring well on the PARCC field test. Second, large numbers of difficult and/or poorly discriminating items may have attenuated the correlation due to restriction of range for much of the high school population of students. Third, it is certainly possible that the larger number of dropped items for high school grades contributed to the lower correlations. The large number of dropped items may have reduced the variance or reliability of the high school PARCC scores. There is also the possibility that the construct being measured by PARCC is more different compared to the state tests for the higher grades. Any or all of these factors, and potentially other factors as well, may help explain the lower correlations for higher grades.

Claim 2: Items are of Sufficient Quality and Rigor[9]

An investigation of test item difficulty demonstrates that many of the PARCC items on the field test were difficult for students. After dropping items for administrative reasons, we examined p-values (proportion of students answering correctly) for test items. ETS provided flagging criteria for p-values to indicate very easy or very difficult items. They were 0.80 and 0.95, to indicate two levels of very easy items, and 0.30 to indicate very difficult items. A well-designed

[9] The term rigor typically refers to a combination of difficulty and cognitive complexity. For this report, only item difficulty was available and investigated. Additional evidence of rigor may come from future alignment studies or other investigations.

assessment will typically have a few very easy and very difficult items and a larger proportion of items that discriminate more toward the center of the distribution. This design ensures that the test is most accurate in the area of the distribution where most students score. PARCC PBA and EOY assessments, for both ELA and math, had very few items flagged as easy (fewer than 2% for all assessments). Many more items were flagged as difficult, especially for the math assessments. For ELA EOY[10], there were between 21% and 47% of items flagged as difficult (fewer than 30% of students answered correctly). More than 50% of items were flagged as difficult for all math grades higher than grade 6 for both PBA and EOY (some subject area tests had as many as 76% of items flagged). Between 18% and 47% of items were flagged as difficult for math grades 3-6. These data are concerning because such a difficult test may not differentiate well among students in the center and lower half of the distribution, limiting the potential interpretations of scores for many students. If the operational assessments exhibit similar item difficulties, it may be prudent to develop more items targeting lower performing students.

In addition to p-values, it is also informative to examine items point biserials, or item-total correlations. These statistics are computed by correlating students' scores on a particular item with their overall total score for the assessment. They tell us if the item functions as expected, or if more able students have a higher likelihood of answering the items correctly than less able students. Negative or near zero correlations typically indicate a problem with the item key or other major issue. We typically drop items with item-total correlations less than 0.20, and we did so for subsequent analyses of PARCC assessments. ELA PBA and EOY assessments had between 2% and 7% of items flagged for item-total correlation. Math grades 3-8 were similar, ranging from 1% to 10% of items flagged. Math course assessments were much large ranging between 9% and 31% of items flagged. This may be due to their being so many very difficult items on the math course assessments attenuating the correlations. We would expect these items to be dropped from operational tests and replaced during item development for future PARCC administrations.

Claim 3: Students Respond to Items as Intended

Survey responses relating to test directions were moderately supportive of the assumption that students understood how to respond to the items as intended. More than 78% of students taking the EOY and PBA, either on the computer or paper-based, claimed that they understood the directions as read to them by the test administrators. About 70% of test administrators agreed that students appeared to understand the instructions read to them. About 60% of test administrators also agreed that the instructions covered all the information necessary to take the test. While these percentages represent a majority of teachers and students supporting

[10] ELA PBA p-values were not calculated due to issues with the test map max score field.

that the instructions were adequate, it leaves a substantial number of field test participants who were not satisfied with the assessment instructions.

Only between 13% and 35% of students responded that they “almost never” found it hard to understand the directions for the questions. Students were much more likely to have difficulty with the directions for questions on math than ELA assessments. Students often claimed that the directions for test questions were unclear or confusing and that they had to re-read the directions several times, which caused delays.

Most students completed a practice exercise prior to taking the PARCC field test assessments, but about 45% of students participating in a computer-based assessment reported that they did not practice prior to testing. Test administrators were more likely to indicate that students had practiced prior to testing, but their results refer to the group of students for whom they administered the test, rather than for individual students. About 75% indicated that their students had practiced for the computer-based assessment. Among students participating in a paper-based assessment, about half of test administrators indicated that their students had not practiced with PARCC sample items prior to testing. Although paper-based assessments won’t require students to gain experience with a new electronic interface, it is concerning that some proportion of these students did not gain any familiarity with PARCC item formats prior to the assessment.

Omit rates

Omit rates are one indicator of whether students understood the instructions for items and responded as intended. If a student does not respond, it can be for several reasons. The item may be confusing or unclear, it may require more effort than other items causing the student to omit it in favor of more easily completed items, it may be toward the end of the test and the student may be fatigued, or other unknown reasons may also come into play. Irrespective of the reason, if a student does not attempt an item, that item does not provide useful information regarding the student’s ability or achievement. This can be particularly troublesome if the omitted items are concentrated within a sub-content area or a particular item type.

The ELA PBA assessments tended to have more omits in the Written Expression and Writing Knowledge Language and Conventions sub-claim. Items under this sub-claim were typically omitted more than twice as often as items from other sub-claims. Omit rates increased as grade level increased, but this pattern among the sub-claims was largely consistent across grades. Omit rates by sub-claim for the ELA PBA ranged from means of 1.44% to 21.88% (means computed by computing omit rates by item and averaging across items within a sub-claim).

ELA EOY omit rates were more consistent by sub-claim, and smaller overall than PBA rates. Mean omit rates ranged from 1.60% to 3.98% by sub-claim. Omit rates did not increase with increasing grade level as with PBA.

Math PBA omit rates tended to increase with increasing grade level as well. Math Fluency items had the lowest omit rates, while items from the Reasoning and Modeling sub-claims tended to have the largest omit rates. Math PBA mean omit rates ranged from 1.05% to 29.27%. Reasoning and Modeling sub-claims for math course tests typically had mean omit rates greater than 20%. This means that 1 of 5 students' responses were omitted for items in these sub-claims. Math EOY results mirror the pattern from ELA, except for Integrated Mathematics, which had much higher omit rates than other math EOY assessments.

When we examined omit rates by item type, it was not surprising that multiple choice items typically had the lowest omit rates. Open-ended items tended to have the highest omit rates, but gridded, multiple part, and standalone technology-enhanced items often had high omit rates as well. As before, course assessments in mathematics exhibited the highest omit rates. Some course tests had mean omit rates for open-ended items greater than 20%.

It is perhaps not surprising that omit rates for open-ended items are larger than for other items. There are no stakes for students taking the PARCC field test, so motivation may be an issue, especially when students are required to write lengthy responses. It is perhaps more troubling that relatively easily completed but more complex items, such as gridded, multiple part, and technology-enhanced items, are also much more likely to be omitted than multiple-choice items. It will be important to monitor this trend during operational testing to determine if the issue is primarily motivation, or if other factors are likely.

Off-topic responses

Much like omitted responses, off-topic responses can also help inform us if students are responding to items as intended. A portion of items from the ELA PBA assessments were hand scored and scorers were given instructions for coding off-topic responses. These responses were outside the content of the item and could indicate frustration or lack of motivation on the part of the student, or perhaps an extreme misunderstanding of the item content. Off topic response rates were low, ranging from about 0.3% to 2% depending on grade. High school students were more than twice as likely to provide off-topic responses than elementary students.

Behavior on multiple select items

Several PARCC items were designed to collect multiple responses or pieces of evidence from students. These items were accompanied by instructions on the specific number of responses students were expected to provide. For example, a student might be asked to provide a response to a multiple choice question in the first portion of the item, and then identify two pieces of evidence in the second portion of that item. Students who respond to the item as intended will select two pieces of evidence. If students misunderstand the question, they might select fewer or more pieces of evidence for the multiple-select item.

The majority of students selected the number of responses indicated in the question (between 63% and 94%). A fairly small proportion of students gave more responses than required (0.1% to 9%), likely indicating that they misunderstood the instructions or did not understand that only the required number of correct responses would be included in the answer choices. A substantial portion of students provided fewer responses than the item required. Many students (as many as 34%) selected only one response for the multiple-select item. It is unknown if the students misunderstood the instructions or did not answer because they could only identify one piece of evidence rather than the required number. The pattern of students identifying one piece of evidence decreases as grade level increases. This may signal that students are unaccustomed to responding to items that require multiple selections in lower grades. We recommend requiring students to participate in the practice tests/tutorials and ensuring that all students see multiple-select items prior to testing. If students do not select the correct number of responses, the tutorial should provide feedback to help orient them to this type of test question.

Person fit

Person fit can tell us if a person’s responses aggregated across items is congruent with a specified IRT model. In the case of the PARCC field test, we used person fit to search for anomalous scoring patterns by student subgroup. We expect a small (about 5%) of respondents to exhibit poor fit, and the PARCC field test assessments followed this pattern. We would be concerned if the students identified for poor person fit were predominantly from a specific subgroup. This was generally not the case. Asian students had relatively high incidence of poor person fit, but not so much as to suggest test bias.

Timing Study

The PARCC assessments were not primarily designed to be speeded tests[11] (tests where speed is considered part of the measurement construct). There are practical limits, however, to how much time may be allowed for students to complete any test. Schools must administer tests within the regular schedule and allocate resources, such as teacher administrators and computers, in reasonable ways.

[11] This statement omits the math fluency sections of the math assessments, which were not analyzed as part of this study.

For several of the PARCC field test assessments, the majority of students were unable to complete the tests during the regularly allotted session time. These students were provided extended time, but for some tests, a substantial proportion did not finish in that amount of time either. HumRRO conducted numerous analyses on student test time and PARCC adjusted session times for operational tests based on those analyses. We recommend that timing studies be repeated for operational tests to ensure that the adjustments were reasonable and that the PARCC assessments are not speeded. As students become more familiar with the content and format of the PARCC assessments, it may become possible to reduce session times.

Math Fluency Study

Some items were designated as math fluency items on the PARCC math tests in grades 3-6. These items were either interspersed on the tests, appeared at the end but students were given no indication that speed was a factor, or appeared at the end of the test and students were told that speed counted towards their scores. Student performance was very similar across conditions. PARCC is not pursuing measuring math fluency with operational tests at this time.

Claim 4: Scores Accurately and Reliably Reflect Student Achievement on the Assessed Content

Scoring site visits

Two HumRRO staff observed the scorer training in June 2014. During this scorer training visit, HumRRO staff separately observed two of three training groups, each comprised of more than 35 experienced scorers. On a second visit in June-July 2014, the same HumRRO staff members separately observed live scoring by two pods. In general, handscoring was consistent with procedures established in the Handscoring Specifications 4.1 document. Scoring directors worked with struggling scorers to help them meet calibration criteria and released scorers who were unable to meet criteria. This ensured that accuracy and consistency levels are maintained throughout scoring. Paper tests and computer tests were tracked separately. We recommend continuing to monitor scoring processes for accuracy and consistency and documenting any differences in scoring statistics between paper and computer administered assessments.

Alternative scoring rules

The PARCC assessments included several two-part items that were scored using a set of scoring rules. Following the PARCC Rules Based Scoring Meeting in Iowa June30-July1, 2014, PARCC requested that certain items be rescored using a set of alternate scoring rules. This included two-part technology enhanced items and evidence-based selected response items with three or more Evidences, as well as two-part technology enhanced items with six or more student responses in the Accuracy portion.

The alternate scoring rules allowed students additional options for scoring 1 or 2 points, where the original scoring rule would have resulted in scores of 0 or 1 point, respectively (the full description of the rules are included in the report). Not surprisingly, the distribution of scores on the items under alternate scoring rules shifted to the right. Because scoring was generally very poor for these items, this shift made the distributions more "normal." In many instances, the overwhelmingly most common score was 0 for these items under the original rule.

We also examined the item-total correlations for these items under the original and alternate scoring rules. The correlations were very similar irrespective of the scoring rules used. We then examined item-total correlations by item part. There was no consistent pattern of either accuracy or evidence having substantially higher item-total correlations. The correlations for some specific item parts were very low. The item parameters tend to yield better information (in the range where most students scored) if the alternate scoring rules were used.

We recommend that PARCC consider the alternate scoring rules for operational testing and that part-level data should be collected and made available for analyses. The overall poor performance of students on these items limits the information that may be assessed from them. Relaxing the scoring criteria results in more normal distributions of student scores. We would also recommend considering scoring the parts separately. Several parts had relatively high item-total correlations independently of their matching part. Alternatively, several parts had very low item-total correlations that could attenuate the results from a matching part that is functioning as intended. This would allow PARCC to evaluate the item parts separately and give them the option of dropping poorly functioning parts while salvaging the item information provided by the matching part. This could increase score reliability.

Conclusion

Conclusions described here refer to the PARCC field test administration. One purpose of the field test was to gather information about items prior to operational administration. A substantial number of items will be dropped or revised prior to operational testing. This is an expected result of field testing and does not represent a criticism of the operational test. The number and proportion of items that do not meet statistical criteria for some grade subjects, especially for math course assessments, may limit the number of forms that can be constructed with adequate content representation.

Claim 1 (Design): The Assessment Connects with Common Core State Standards (CCSS)

Once we eliminated administratively dropped and poorly functioning items, we examined correlations with state assessment results. Correlations were generally strong and followed expected patterns for convergent/discriminant validity coefficients for elementary and middle school grades. High school grade correlations were weaker, but still followed the generally expected pattern. This represents one piece of evidence supporting the link between the PARCC assessments and the CCSS. Follow up studies should be conducted to verify that the PARCC assessments adequately connect with CCSS, and particular attention should be given to high school mathematics assessments.

Claim 2 (Development): Items are of Sufficient Quality and Rigor

More items than expected were dropped based on administration and statistical quality (item-total correlation). PARCC may need to develop more items than originally expected during the early years of administration to generate the anticipated number of test forms. No further analyses were conducted on items that were dropped.

Item difficulty, which we will use as a proxy for test rigor, was higher than expected. Many items were flagged for very low p-values (few students getting the correct response). This was especially true of math course assessments. Consequently, we recommend focusing item development to ensure that a substantive number of less difficult items are included. We also recommend close monitoring of student classification distributions, overall and by subgroup, and classification accuracy to ensure robust measurement throughout the test scale.

Claim 3 (Administration): Students Respond to Items as Intended

A substantial proportion of students and administrators described issues with the overall test instructions or with item level instructions. Omit rates were high for complex and high-effort item types, especially for high school level students. A substantial proportion of students did not respond as expected to multiple select items, often selecting only one response when asked to select two or more. Some students described confusion related to use of test tools and test navigation. These are things we hope to discover during field test, but they should be surveyed again during operational testing to ensure that these issues have been adequately addressed.

Field test session times were not adequate for several of the sessions. Testing session times were revised based on field-test results, but should be monitored to ensure that students have adequate time to complete each session and to guard against speeded conditions.

Claim 4 (Scoring): Scores Accurately and Reliably Reflect Student Achievement on the Assessed Content

Items requiring handscoring were scored according to PARCC guidelines. Scorers were regularly monitored for consistency and accuracy and retrained or dismissed if they did not reach acceptable levels. We recommend continuing these processes for operational testing and conducting rater drift analyses in the future to ensure consistent scoring within and across years.

Two-part technology enhanced items and evidence-based selected response items with three or more Evidences, as well as two-part technology enhanced items with six or more student responses in the Accuracy portion, were rescored using alternate scoring rules. The alternate scoring rules yielded more normal distributions of student scores, compared to the original rules, which yielded distributions shifted to the left (typically with many 0 scores). The alternate scoring rules did not consistently impact item-total correlations. Investigations of "part-level" scores indicate that item parts may be contributing information on their own. In some instances, one part would function well, but the other poorly, attenuating the information that could be obtained for the item. We recommend that PARCC consider the alternate scoring rules to shift the item-level score distributions toward normal. We also recommend that PARCC consider scoring the item parts separately to optimize test information and to allow for the omission of a poorly performing part, without losing the entire item.

Cautions

The studies described in this report represent HumRRO's investigations. ETS is also conducting numerous psychometric evaluations of the PARCC field test items. HumRRO and ETS strived to avoid duplication of effort. This report and ETS' report related to the psychometric properties of PARCC field test items should be considered in tandem.

References

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

Drasgow, F., Levine, M.V. & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*, 171-191.

Kane, M.T. (2006). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319-342.

Li, M. F., and Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21*, 215-231.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14*, 127-137.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person fit in IRT models. *Multivariate Behavioral Research, 35*, 543–568.

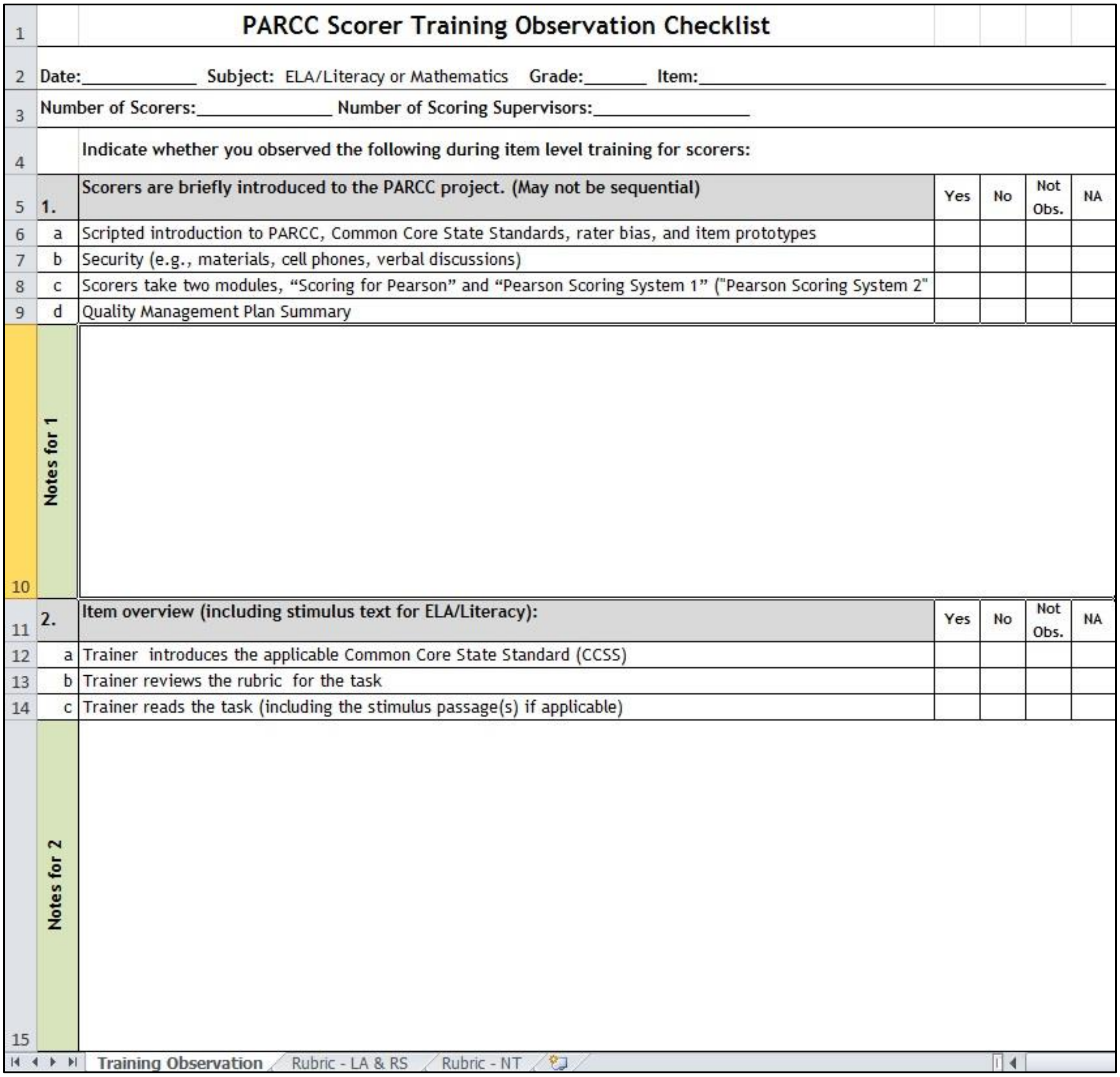
Seo, D. G., & Weiss, D. J. (2013). l_z Person-fit index to identify misfit students with achievement test data. *Educational and Psychological Measurement, 73*, 994-1016.

Sinclair, Deatz, Johnston-Fisher, Levinson, & Thacker (February 2, 2015). *Findings from the quality of test administration investigations: PARCC field test* (HumRRO Report 2014 No. 050). Alexandria, VA: Human Resources Research Organization.

Thacker, A., Sinclair, A., Wise, L., & Becker, S. (April 10, 2014). *PARCC validity studies including predictive and longitudinal studies memorandum* (HumRRO Report 2014 No. 020). Alexandria, VA: Human Resources Research Organization.

Appendix A. Final Scoring Observation Checklists

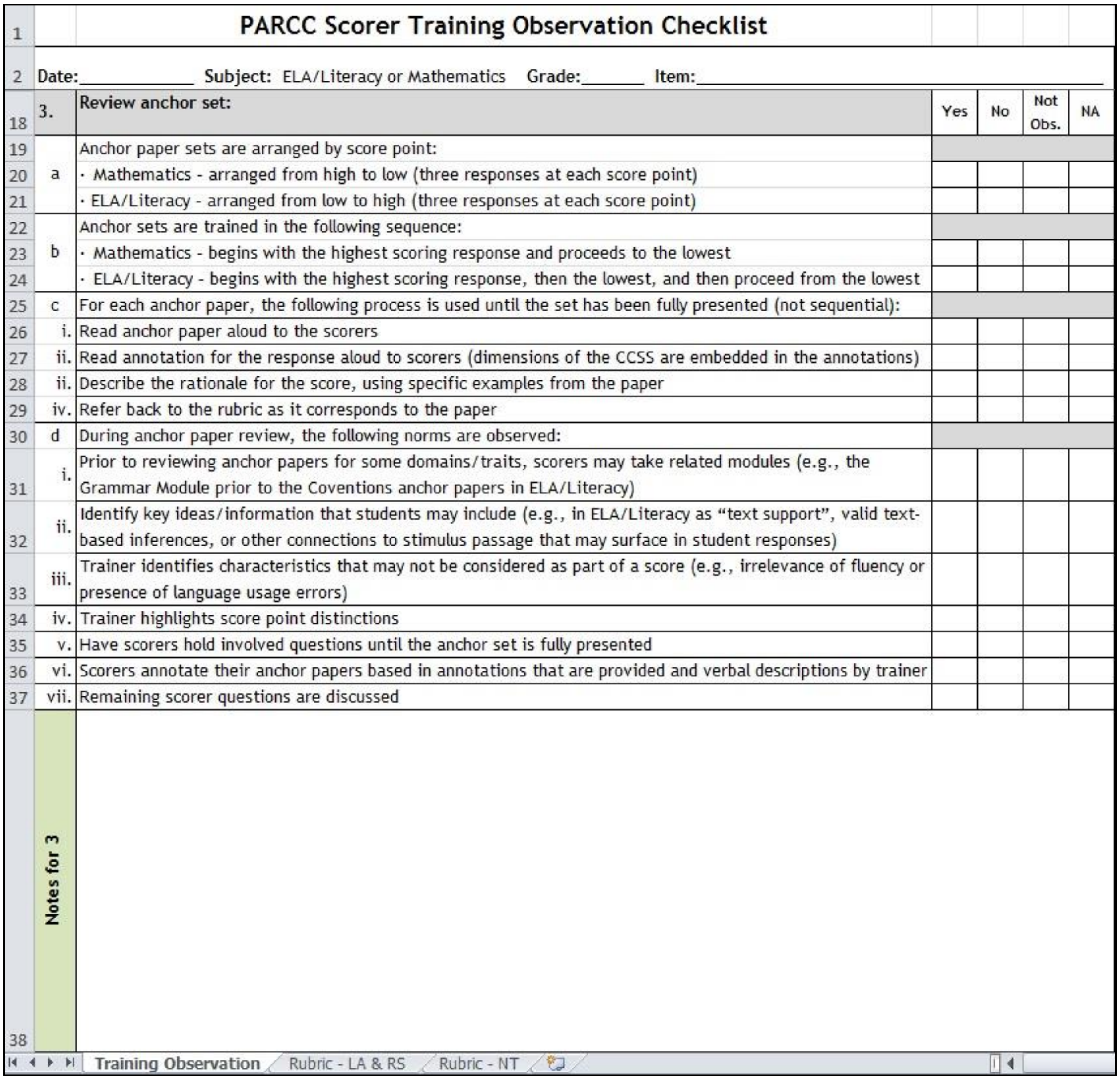
Table A-1. Observation Checklist Used for Scorer Training Observations



1		**PARCC Scorer Training Observation Checklist**				
2	Date:____________	Subject: ELA/Literacy or Mathematics Grade:_______ Item:_______________________				
3	Number of Scorers:______________	Number of Scoring Supervisors:________________				
4		Indicate whether you observed the following during item level training for scorers:				
5	1.	Scorers are briefly introduced to the PARCC project. (May not be sequential)	Yes	No	Not Obs.	NA
6	a	Scripted introduction to PARCC, Common Core State Standards, rater bias, and item prototypes				
7	b	Security (e.g., materials, cell phones, verbal discussions)				
8	c	Scorers take two modules, “Scoring for Pearson” and “Pearson Scoring System 1” ("Pearson Scoring System 2"				
9	d	Quality Management Plan Summary				
10	Notes for 1					
11	2.	Item overview (including stimulus text for ELA/Literacy):	Yes	No	Not Obs.	NA
12	a	Trainer introduces the applicable Common Core State Standard (CCSS)				
13	b	Trainer reviews the rubric for the task				
14	c	Trainer reads the task (including the stimulus passage(s) if applicable)				
15	Notes for 2					

Scorer Training Page 1

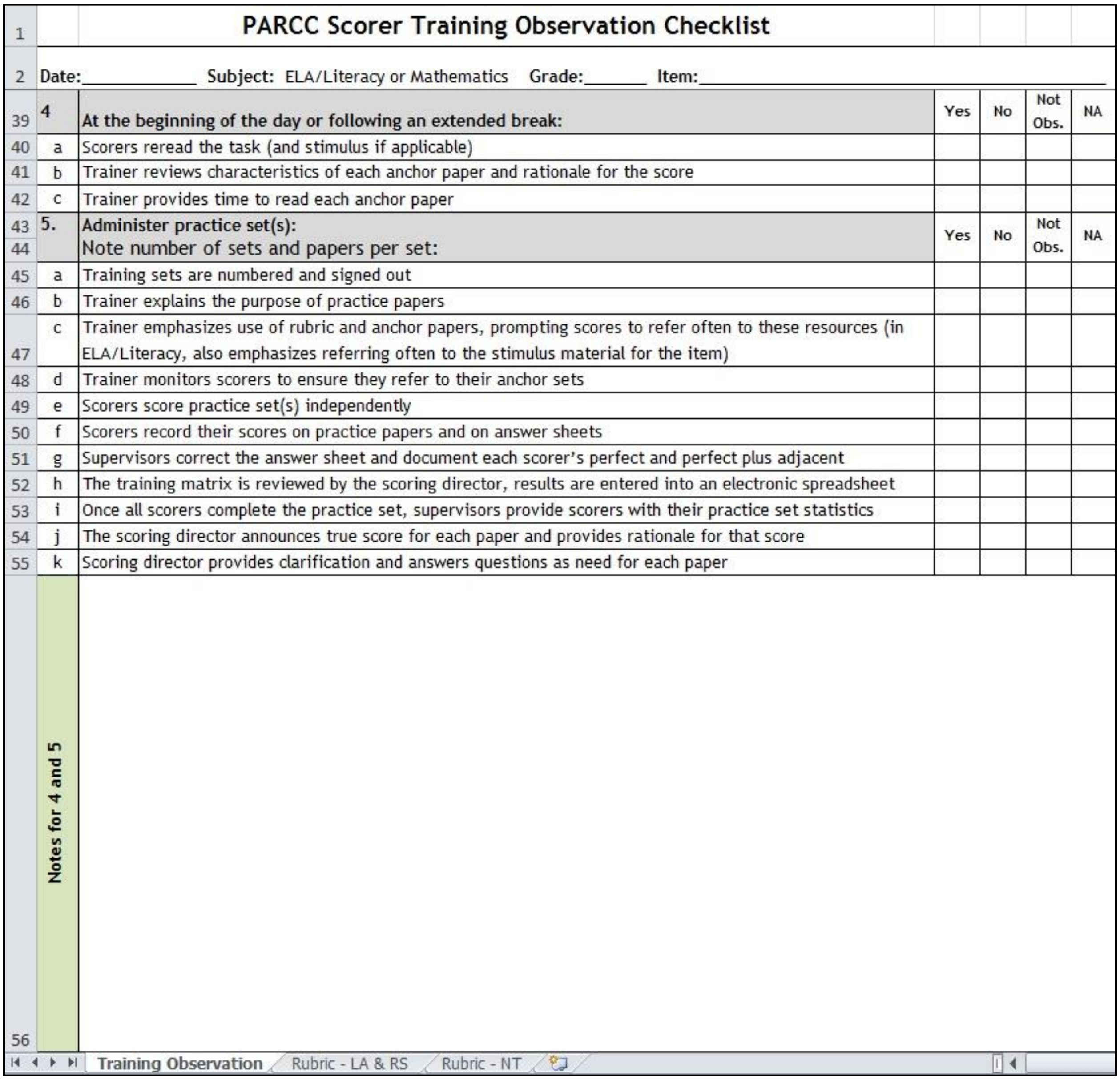
Observation Checklist Used for Scorer Training Observations (continued)



1		PARCC Scorer Training Observation Checklist				
2	Date:____________ Subject: ELA/Literacy or Mathematics Grade:_______ Item:_______________					
18	3.	Review anchor set:	Yes	No	Not Obs.	NA
19	a	Anchor paper sets are arranged by score point:				
20		· Mathematics - arranged from high to low (three responses at each score point)				
21		· ELA/Literacy - arranged from low to high (three responses at each score point)				
22	b	Anchor sets are trained in the following sequence:				
23		· Mathematics - begins with the highest scoring response and proceeds to the lowest				
24		· ELA/Literacy - begins with the highest scoring response, then the lowest, and then proceed from the lowest				
25	c	For each anchor paper, the following process is used until the set has been fully presented (not sequential):				
26	i.	Read anchor paper aloud to the scorers				
27	ii.	Read annotation for the response aloud to scorers (dimensions of the CCSS are embedded in the annotations)				
28	ii.	Describe the rationale for the score, using specific examples from the paper				
29	iv.	Refer back to the rubric as it corresponds to the paper				
30	d	During anchor paper review, the following norms are observed:				
31	i.	Prior to reviewing anchor papers for some domains/traits, scorers may take related modules (e.g., the Grammar Module prior to the Coventions anchor papers in ELA/Literacy)				
32	ii.	Identify key ideas/information that students may include (e.g., in ELA/Literacy as “text support”, valid text-based inferences, or other connections to stimulus passage that may surface in student responses)				
33	iii.	Trainer identifies characteristics that may not be considered as part of a score (e.g., irrelevance of fluency or presence of language usage errors)				
34	iv.	Trainer highlights score point distinctions				
35	v.	Have scorers hold involved questions until the anchor set is fully presented				
36	vi.	Scorers annotate their anchor papers based in annotations that are provided and verbal descriptions by trainer				
37	vii.	Remaining scorer questions are discussed				
38	Notes for 3					

Scorer Training Page 2

Observation Checklist Used for Scorer Training Observations (continued)



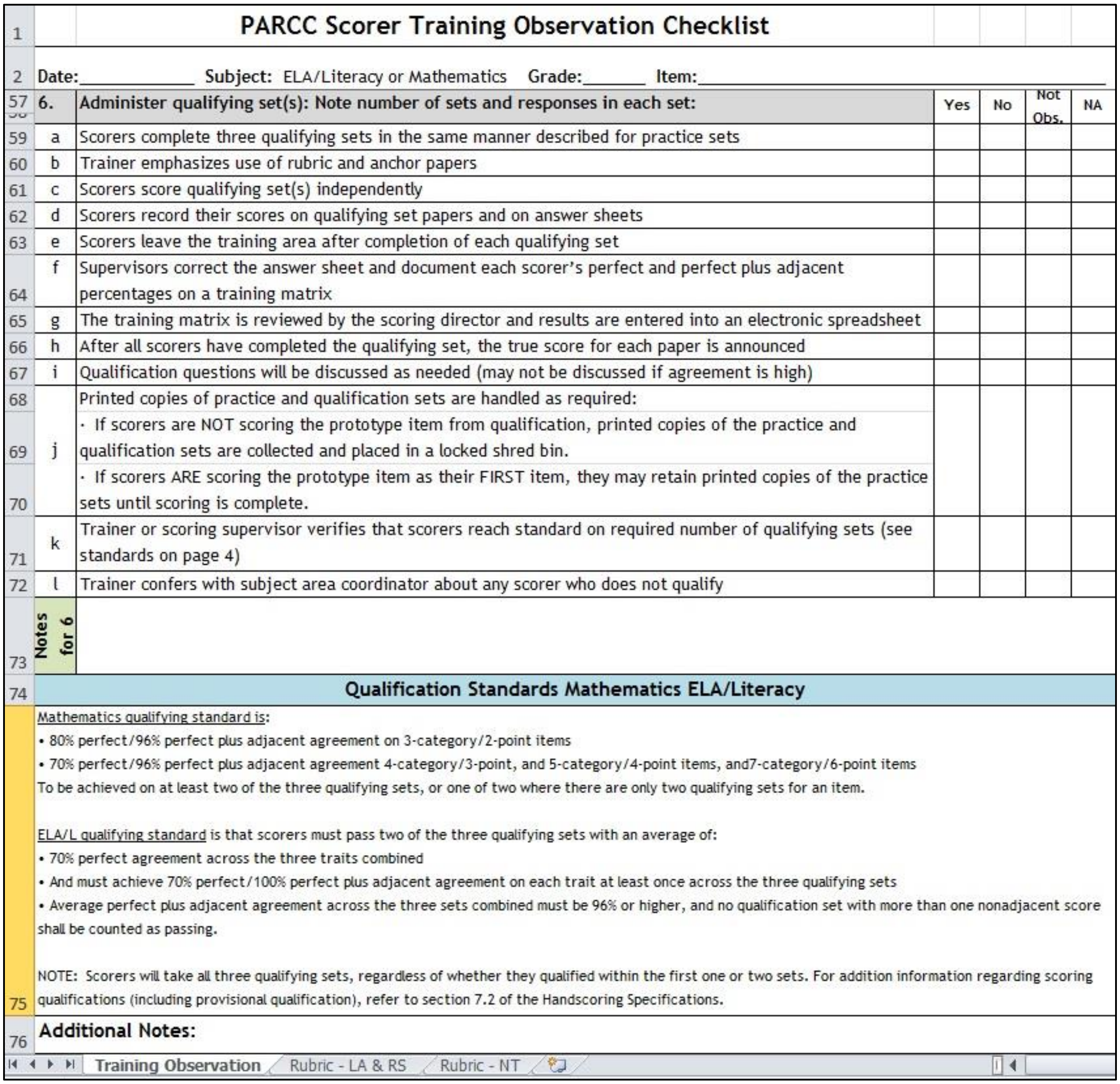
PARCC Scorer Training Observation Checklist

Date:____________ Subject: ELA/Literacy or Mathematics Grade:_______ Item:____

4	**At the beginning of the day or following an extended break:**	Yes	No	Not Obs.	NA
a	Scorers reread the task (and stimulus if applicable)				
b	Trainer reviews characteristics of each anchor paper and rationale for the score				
c	Trainer provides time to read each anchor paper				
5.	**Administer practice set(s):** Note number of sets and papers per set:	Yes	No	Not Obs.	NA
a	Training sets are numbered and signed out				
b	Trainer explains the purpose of practice papers				
c	Trainer emphasizes use of rubric and anchor papers, prompting scores to refer often to these resources (in ELA/Literacy, also emphasizes referring often to the stimulus material for the item)				
d	Trainer monitors scorers to ensure they refer to their anchor sets				
e	Scorers score practice set(s) independently				
f	Scorers record their scores on practice papers and on answer sheets				
g	Supervisors correct the answer sheet and document each scorer's perfect and perfect plus adjacent				
h	The training matrix is reviewed by the scoring director, results are entered into an electronic spreadsheet				
i	Once all scorers complete the practice set, supervisors provide scorers with their practice set statistics				
j	The scoring director announces true score for each paper and provides rationale for that score				
k	Scoring director provides clarification and answers questions as need for each paper				
Notes for 4 and 5					

Scorer Training Page 3

Observation Checklist Used for Scorer Training Observations (continued)



1		**PARCC Scorer Training Observation Checklist**				
2	Date:____________ Subject: ELA/Literacy or Mathematics Grade:_______ Item:______________________					
57	**6.**	**Administer qualifying set(s): Note number of sets and responses in each set:**	Yes	No	Not Obs.	NA
59	a	Scorers complete three qualifying sets in the same manner described for practice sets				
60	b	Trainer emphasizes use of rubric and anchor papers				
61	c	Scorers score qualifying set(s) independently				
62	d	Scorers record their scores on qualifying set papers and on answer sheets				
63	e	Scorers leave the training area after completion of each qualifying set				
64	f	Supervisors correct the answer sheet and document each scorer's perfect and perfect plus adjacent percentages on a training matrix				
65	g	The training matrix is reviewed by the scoring director and results are entered into an electronic spreadsheet				
66	h	After all scorers have completed the qualifying set, the true score for each paper is announced				
67	i	Qualification questions will be discussed as needed (may not be discussed if agreement is high)				
68	j	Printed copies of practice and qualification sets are handled as required:				
69		· If scorers are NOT scoring the prototype item from qualification, printed copies of the practice and qualification sets are collected and placed in a locked shred bin.				
70		· If scorers ARE scoring the prototype item as their FIRST item, they may retain printed copies of the practice sets until scoring is complete.				
71	k	Trainer or scoring supervisor verifies that scorers reach standard on required number of qualifying sets (see standards on page 4)				
72	l	Trainer confers with subject area coordinator about any scorer who does not qualify				
73	**Notes for 6**					
74	**Qualification Standards Mathematics ELA/Literacy**					

75

Mathematics qualifying standard is:

- 80% perfect/96% perfect plus adjacent agreement on 3-category/2-point items
- 70% perfect/96% perfect plus adjacent agreement 4-category/3-point, and 5-category/4-point items, and7-category/6-point items

To be achieved on at least two of the three qualifying sets, or one of two where there are only two qualifying sets for an item.

ELA/L qualifying standard is that scorers must pass two of the three qualifying sets with an average of:

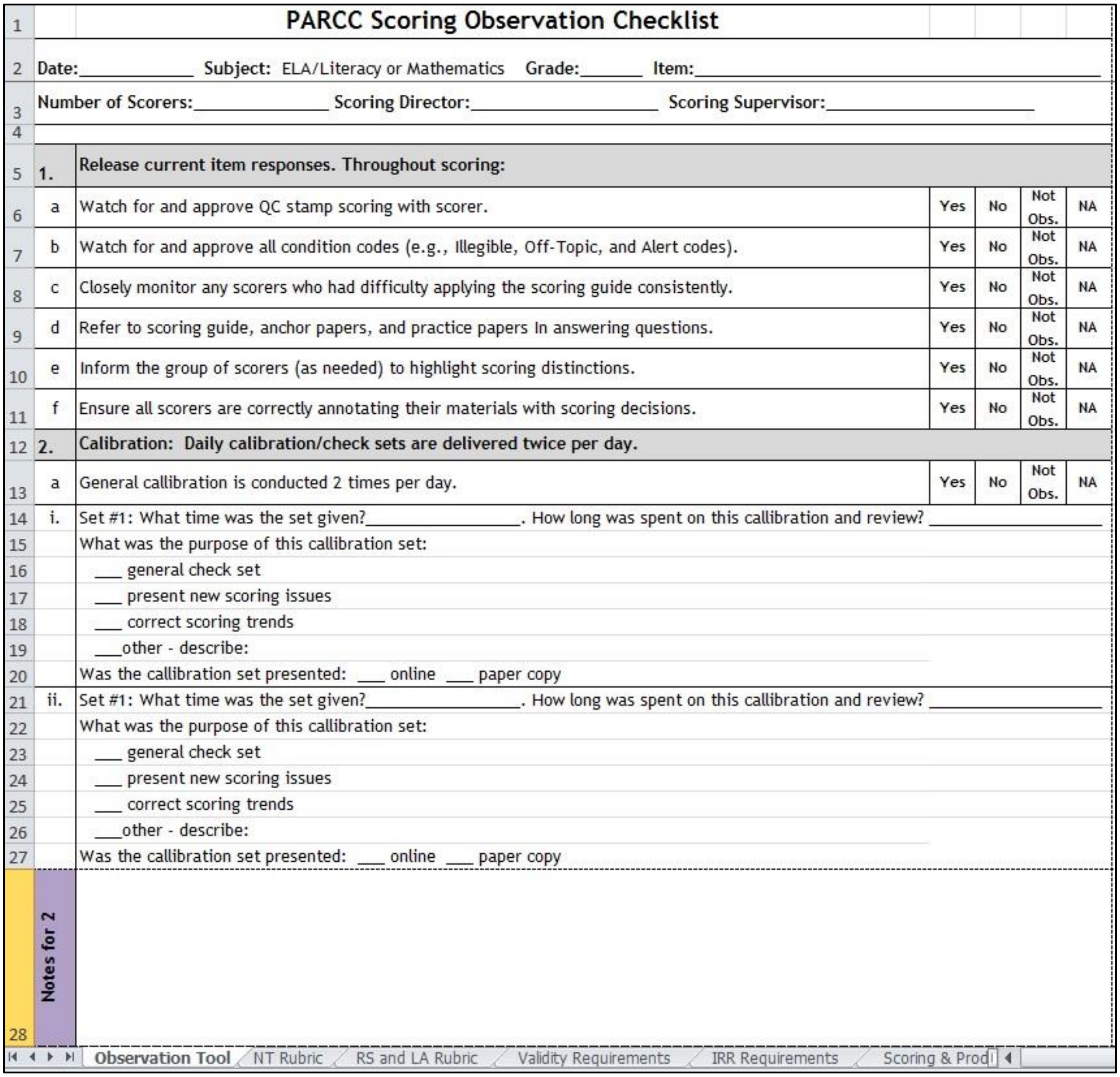
- 70% perfect agreement across the three traits combined
- And must achieve 70% perfect/100% perfect plus adjacent agreement on each trait at least once across the three qualifying sets
- Average perfect plus adjacent agreement across the three sets combined must be 96% or higher, and no qualification set with more than one nonadjacent score shall be counted as passing.

NOTE: Scorers will take all three qualifying sets, regardless of whether they qualified within the first one or two sets. For addition information regarding scoring qualifications (including provisional qualification), refer to section 7.2 of the Handscoring Specifications.

76 **Additional Notes:**

Scorer Training Page 4

Table A-2. Observation Checklist Used for Item Scoring Observation



PARCC Scoring Observation Checklist

Date:____________ Subject: ELA/Literacy or Mathematics Grade:_______ Item:____

Number of Scorers:______________ Scoring Director:____________________ Scoring Supervisor:______________________

1.	**Release current item responses. Throughout scoring:**				
a	Watch for and approve QC stamp scoring with scorer.	Yes	No	Not Obs.	NA
b	Watch for and approve all condition codes (e.g., Illegible, Off-Topic, and Alert codes).	Yes	No	Not Obs.	NA
c	Closely monitor any scorers who had difficulty applying the scoring guide consistently.	Yes	No	Not Obs.	NA
d	Refer to scoring guide, anchor papers, and practice papers In answering questions.	Yes	No	Not Obs.	NA
e	Inform the group of scorers (as needed) to highlight scoring distinctions.	Yes	No	Not Obs.	NA
f	Ensure all scorers are correctly annotating their materials with scoring decisions.	Yes	No	Not Obs.	NA
2.	**Calibration: Daily calibration/check sets are delivered twice per day.**				
a	General callibration is conducted 2 times per day.	Yes	No	Not Obs.	NA

i. Set #1: What time was the set given?________________. How long was spent on this callibration and review? __________________

What was the purpose of this callibration set:

___ general check set

___ present new scoring issues

___ correct scoring trends

___other - describe:

Was the callibration set presented: ___ online ___ paper copy

ii. Set #1: What time was the set given?________________. How long was spent on this callibration and review? __________________

What was the purpose of this callibration set:

___ general check set

___ present new scoring issues

___ correct scoring trends

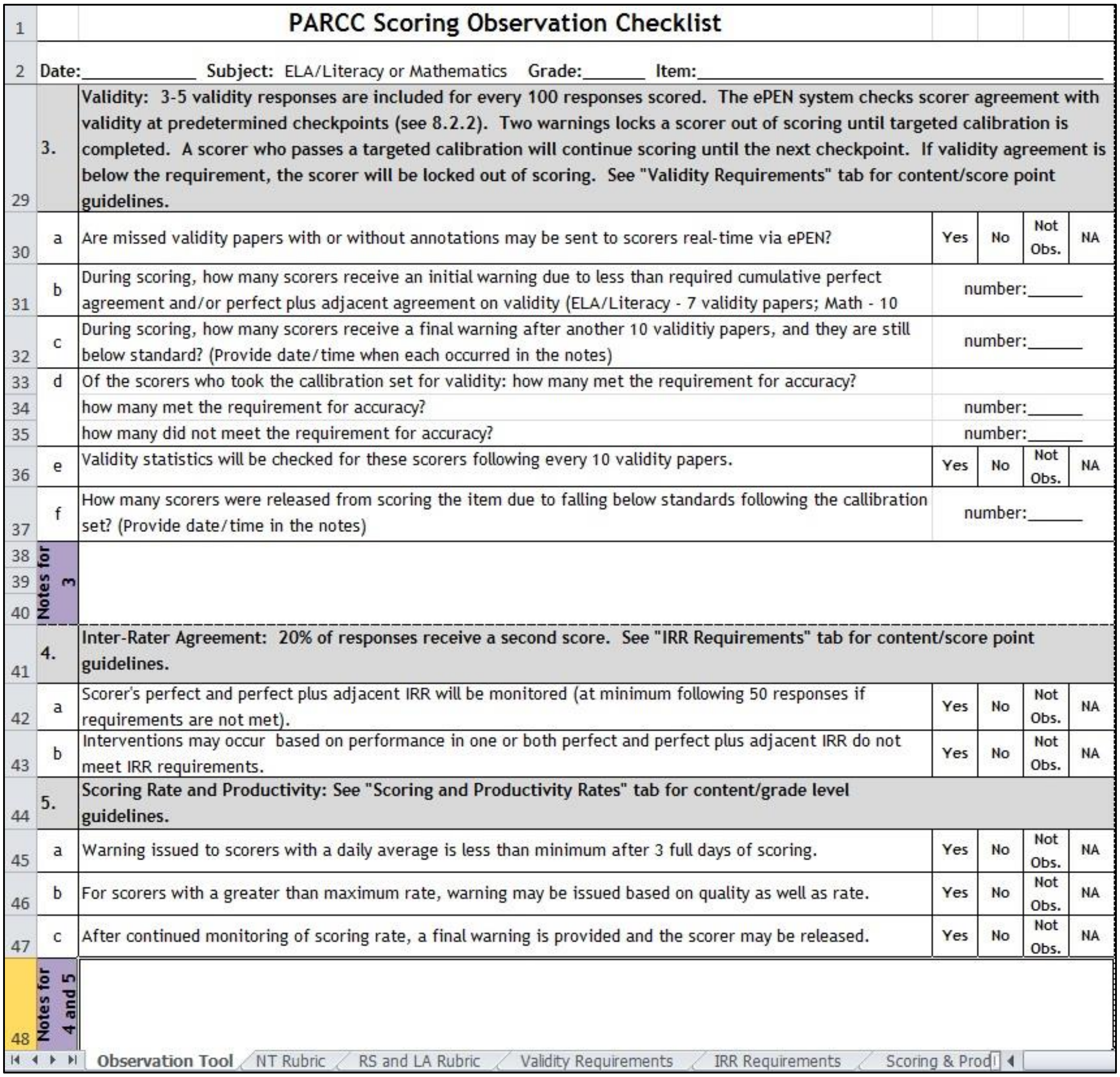
___other - describe:

Was the callibration set presented: ___ online ___ paper copy

Notes for 2

Item Scoring Page 1

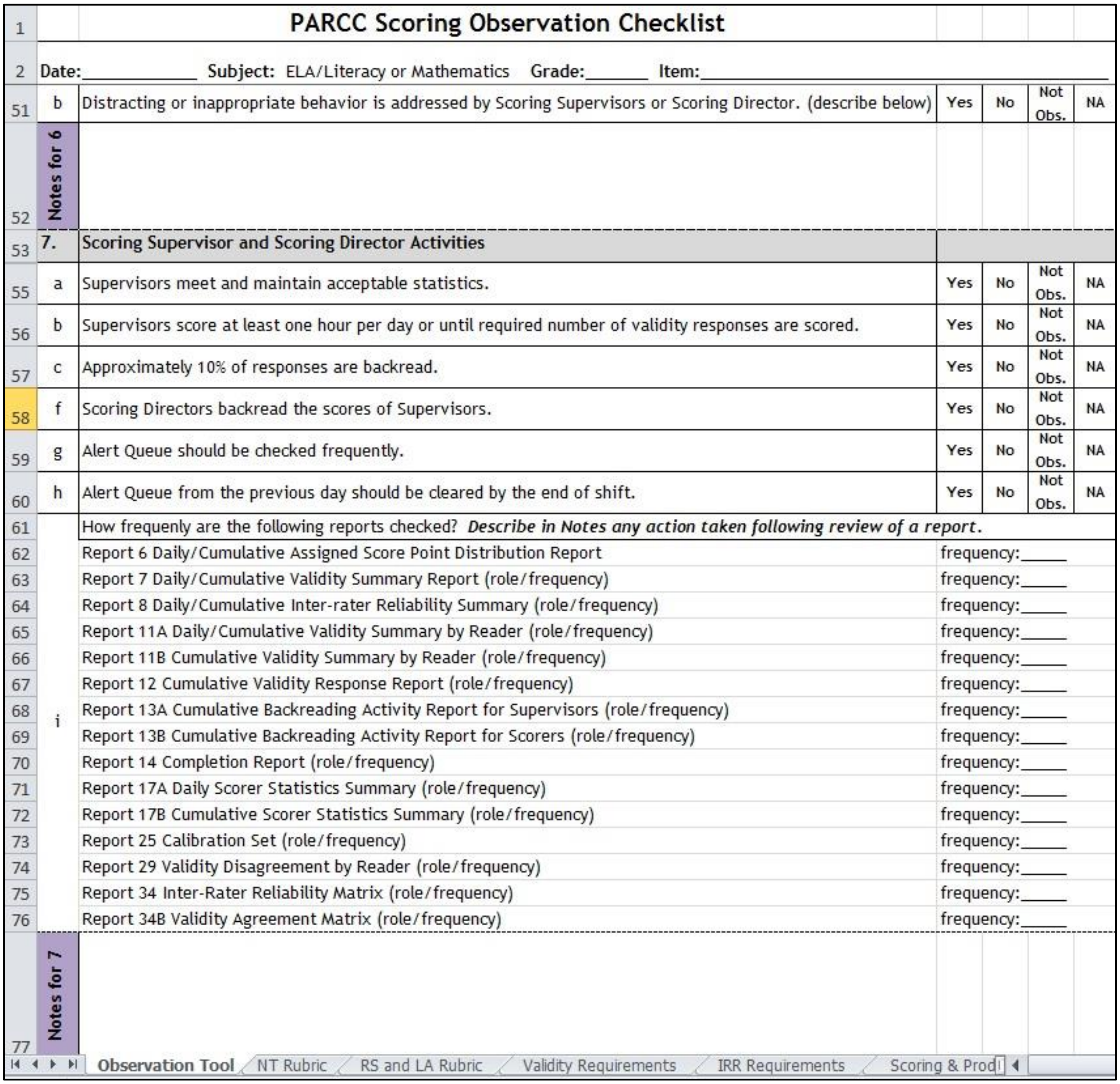
Observation Checklist Used for Item Scoring Observation (continued)



1		**PARCC Scoring Observation Checklist**				
2	Date:____________	Subject: ELA/Literacy or Mathematics Grade:_______ Item:______________________________				
29	3.	**Validity: 3-5 validity responses are included for every 100 responses scored. The ePEN system checks scorer agreement with validity at predetermined checkpoints (see 8.2.2). Two warnings locks a scorer out of scoring until targeted calibration is completed. A scorer who passes a targeted calibration will continue scoring until the next checkpoint. If validity agreement is below the requirement, the scorer will be locked out of scoring. See "Validity Requirements" tab for content/score point guidelines.**				
30	a	Are missed validity papers with or without annotations may be sent to scorers real-time via ePEN?	Yes	No	Not Obs.	NA
31	b	During scoring, how many scorers receive an initial warning due to less than required cumulative perfect agreement and/or perfect plus adjacent agreement on validity (ELA/Literacy - 7 validity papers; Math - 10	number:______			
32	c	During scoring, how many scorers receive a final warning after another 10 validitiy papers, and they are still below standard? (Provide date/time when each occurred in the notes)	number:______			
33	d	Of the scorers who took the callibration set for validity: how many met the requirement for accuracy?				
34		how many met the requirement for accuracy?	number:______			
35		how many did not meet the requirement for accuracy?	number:______			
36	e	Validity statistics will be checked for these scorers following every 10 validity papers.	Yes	No	Not Obs.	NA
37	f	How many scorers were released from scoring the item due to falling below standards following the callibration set? (Provide date/time in the notes)	number:______			
38	Notes for 3					
39						
40						
41	4.	**Inter-Rater Agreement: 20% of responses receive a second score. See "IRR Requirements" tab for content/score point guidelines.**				
42	a	Scorer's perfect and perfect plus adjacent IRR will be monitored (at minimum following 50 responses if requirements are not met).	Yes	No	Not Obs.	NA
43	b	Interventions may occur based on performance in one or both perfect and perfect plus adjacent IRR do not meet IRR requirements.	Yes	No	Not Obs.	NA
44	5.	**Scoring Rate and Productivity: See "Scoring and Productivity Rates" tab for content/grade level guidelines.**				
45	a	Warning issued to scorers with a daily average is less than minimum after 3 full days of scoring.	Yes	No	Not Obs.	NA
46	b	For scorers with a greater than maximum rate, warning may be issued based on quality as well as rate.	Yes	No	Not Obs.	NA
47	c	After continued monitoring of scoring rate, a final warning is provided and the scorer may be released.	Yes	No	Not Obs.	NA
48	Notes for 4 and 5					

Item Scoring Page 2

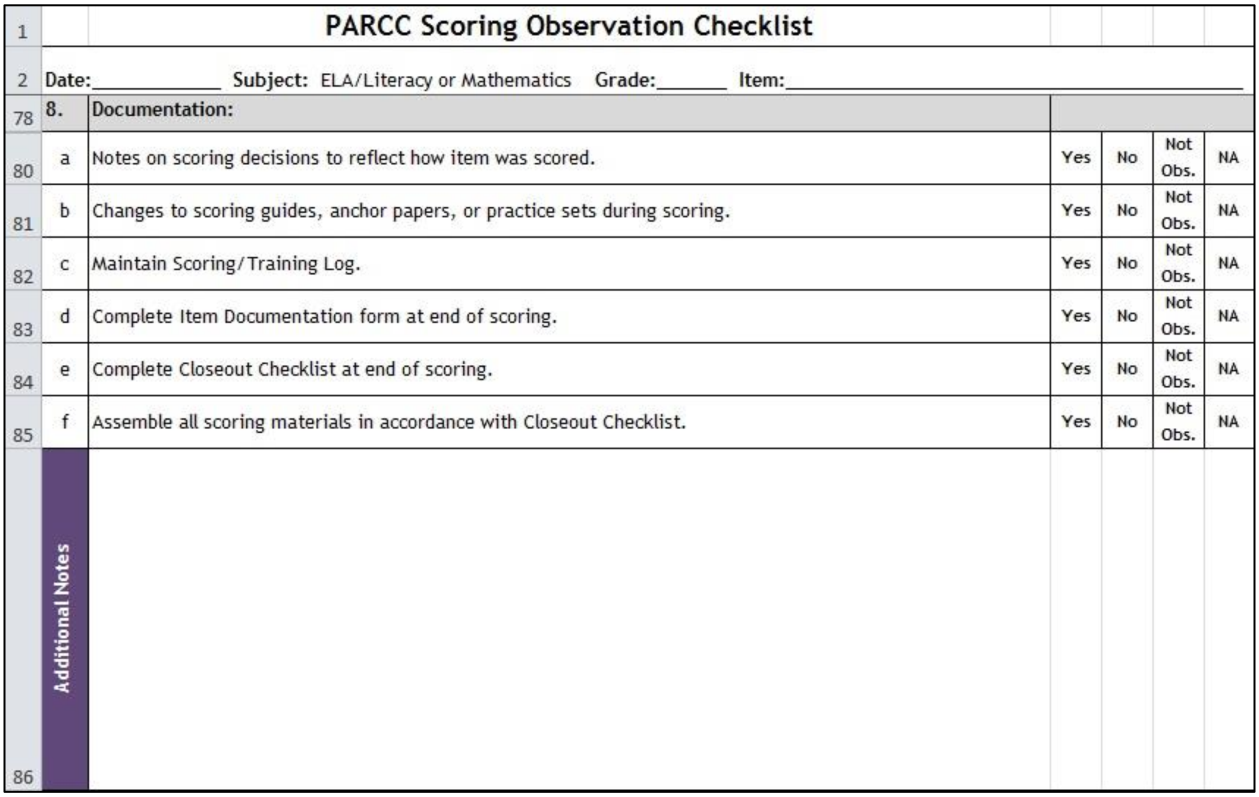
Observation Checklist Used for Item Scoring Observation (continued)



1		**PARCC Scoring Observation Checklist**				
2	**Date:**____________ **Subject:** ELA/Literacy or Mathematics **Grade:**_______ **Item:**________________					
51	b	Distracting or inappropriate behavior is addressed by Scoring Supervisors or Scoring Director. (describe below)	Yes	No	Not Obs.	NA
52	**Notes for 6**					
53	**7.**	**Scoring Supervisor and Scoring Director Activities**				
55	a	Supervisors meet and maintain acceptable statistics.	Yes	No	Not Obs.	NA
56	b	Supervisors score at least one hour per day or until required number of validity responses are scored.	Yes	No	Not Obs.	NA
57	c	Approximately 10% of responses are backread.	Yes	No	Not Obs.	NA
58	f	Scoring Directors backread the scores of Supervisors.	Yes	No	Not Obs.	NA
59	g	Alert Queue should be checked frequently.	Yes	No	Not Obs.	NA
60	h	Alert Queue from the previous day should be cleared by the end of shift.	Yes	No	Not Obs.	NA
61	i	How frequenly are the following reports checked? ***Describe in Notes any action taken following review of a report.***				
62		Report 6 Daily/Cumulative Assigned Score Point Distribution Report	frequency:_____			
63		Report 7 Daily/Cumulative Validity Summary Report (role/frequency)	frequency:_____			
64		Report 8 Daily/Cumulative Inter-rater Reliability Summary (role/frequency)	frequency:_____			
65		Report 11A Daily/Cumulative Validity Summary by Reader (role/frequency)	frequency:_____			
66		Report 11B Cumulative Validity Summary by Reader (role/frequency)	frequency:_____			
67		Report 12 Cumulative Validity Response Report (role/frequency)	frequency:_____			
68		Report 13A Cumulative Backreading Activity Report for Supervisors (role/frequency)	frequency:_____			
69		Report 13B Cumulative Backreading Activity Report for Scorers (role/frequency)	frequency:_____			
70		Report 14 Completion Report (role/frequency)	frequency:_____			
71		Report 17A Daily Scorer Statistics Summary (role/frequency)	frequency:_____			
72		Report 17B Cumulative Scorer Statistics Summary (role/frequency)	frequency:_____			
73		Report 25 Calibration Set (role/frequency)	frequency:_____			
74		Report 29 Validity Disagreement by Reader (role/frequency)	frequency:_____			
75		Report 34 Inter-Rater Reliability Matrix (role/frequency)	frequency:_____			
76		Report 34B Validity Agreement Matrix (role/frequency)	frequency:_____			
77	**Notes for 7**					

Item Scoring Page 3

Observation Checklist Used for Item Scoring Observation (continued)



1		**PARCC Scoring Observation Checklist**				
2	**Date:**____________ **Subject:** ELA/Literacy or Mathematics **Grade:**_______ **Item:**____					
78	**8.**	**Documentation:**				
80	a	Notes on scoring decisions to reflect how item was scored.	Yes	No	Not Obs.	NA
81	b	Changes to scoring guides, anchor papers, or practice sets during scoring.	Yes	No	Not Obs.	NA
82	c	Maintain Scoring/Training Log.	Yes	No	Not Obs.	NA
83	d	Complete Item Documentation form at end of scoring.	Yes	No	Not Obs.	NA
84	e	Complete Closeout Checklist at end of scoring.	Yes	No	Not Obs.	NA
85	f	Assemble all scoring materials in accordance with Closeout Checklist.	Yes	No	Not Obs.	NA
86	**Additional Notes**					

Item Scoring Page 4

Appendix B. Distribution of Person Fit Statistics

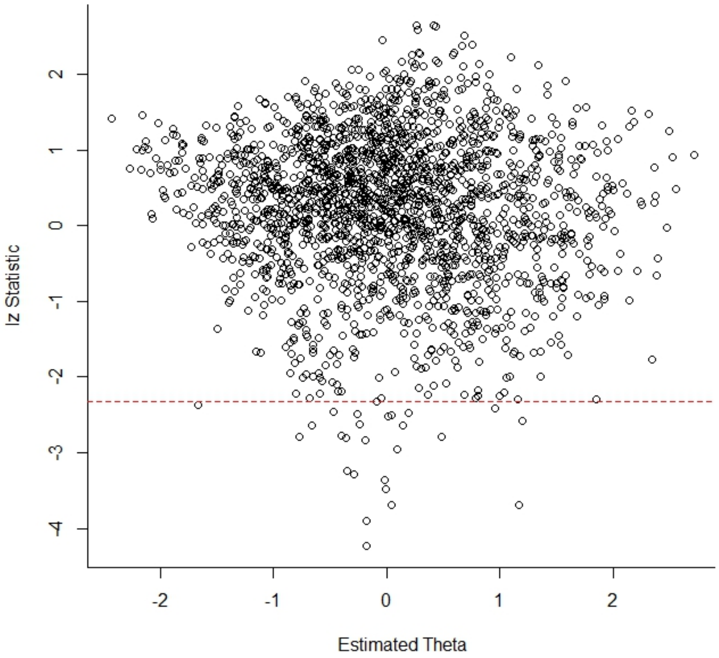


Figure B-1. Person fit distribution of the performance-based ELA Grade 3 assessment, Form 184PO. Scores below the red line indicate poor fit.

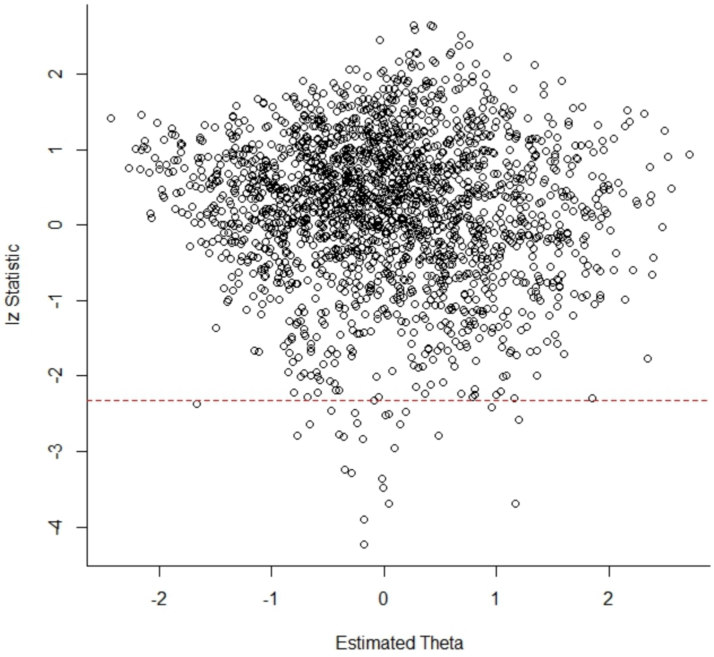


Figure B-2. Person fit distribution of the performance-based ELA Grade 5 assessment, Form 214PO. Scores below the red line indicate poor fit.

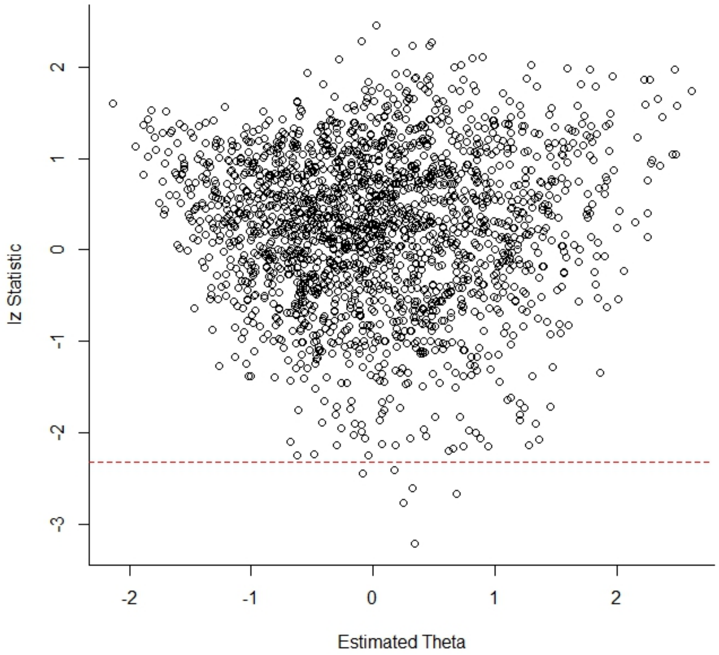


Figure B-3. Person fit distribution of the performance-based ELA Grade 7 assessment, Form 074PP. Scores below the red line indicate poor fit.

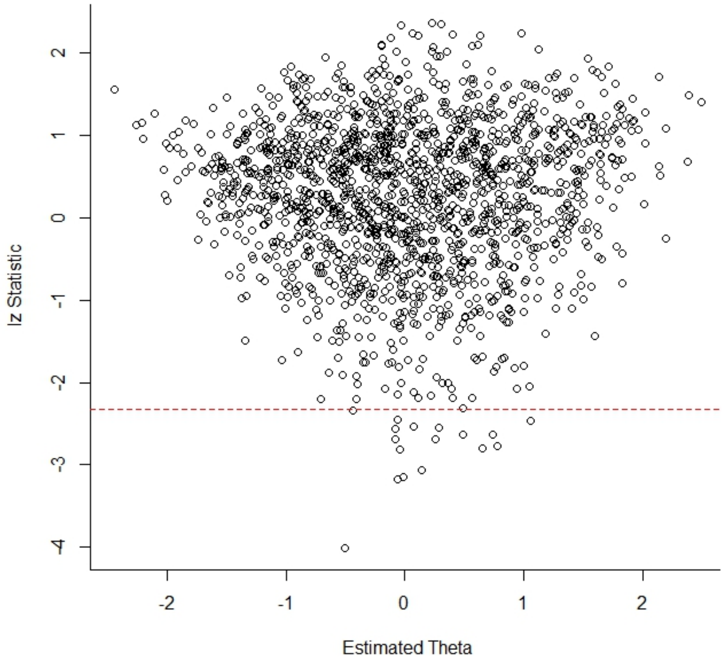


Figure B-4. Person fit distribution of the performance-based ELA Grade 9 assessment, Form 184PO. Scores below the red line indicate poor fit.

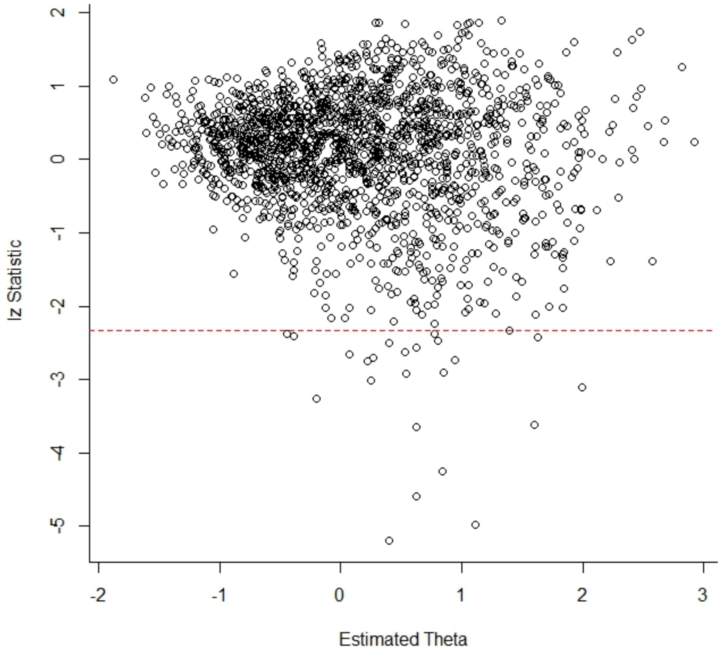


Figure B-5. Person fit distribution of the performance-based Math Grade 3 assessment, Form 014PO. Scores below the red line indicate poor fit.

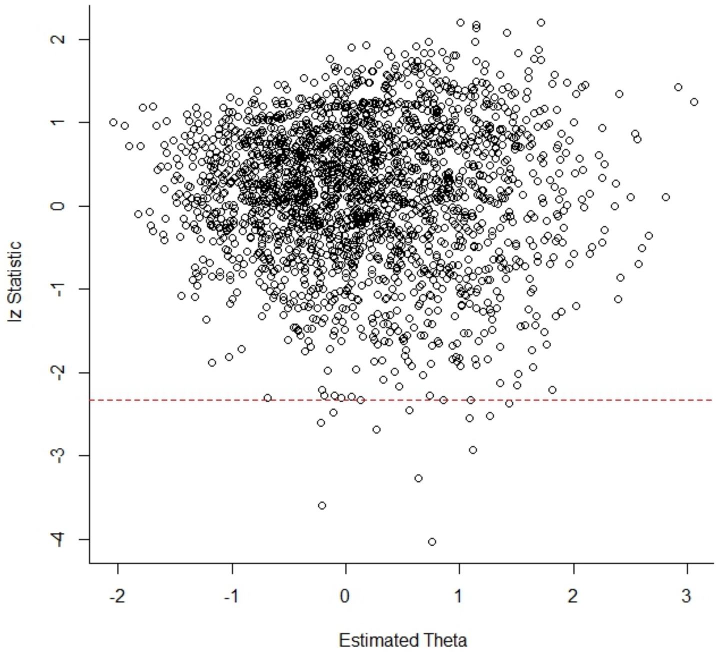


Figure B-6. Person fit distribution of the performance-based Math Grade 5 assessment, Form 124PO. Scores below the red line indicate poor fit.

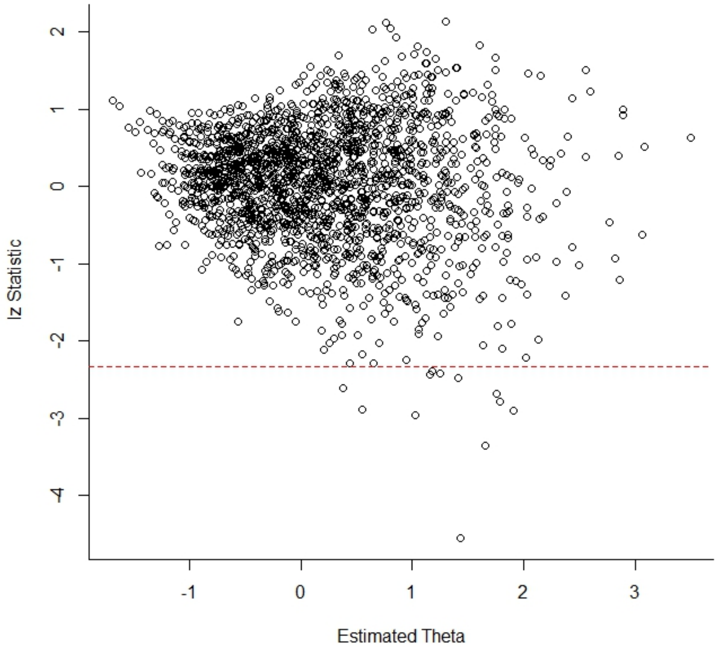


Figure B-7. Person fit distribution of the performance-based Math Grade 7 assessment, Form 114PO. Scores below the red line indicate poor fit.

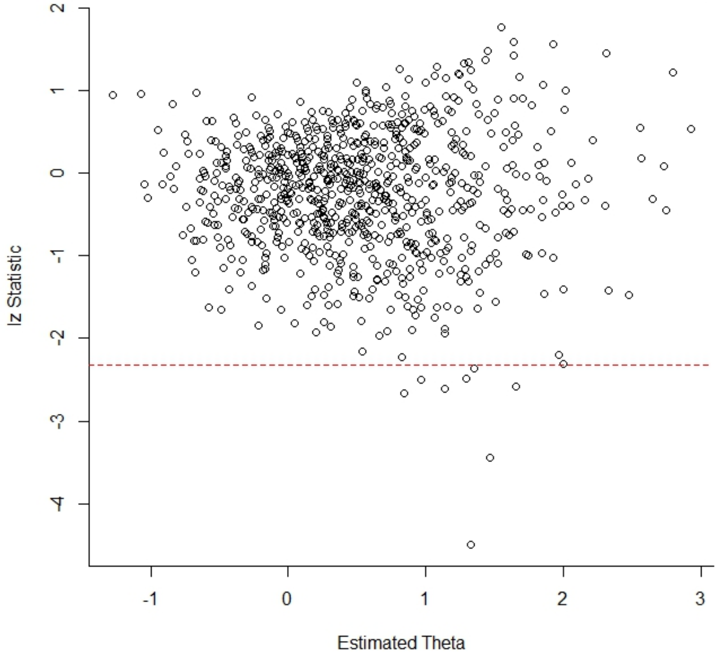


Figure B-8. Person fit distribution of the performance-based Algebra 1 assessment, Form 124PO. Scores below the red line indicate poor fit.

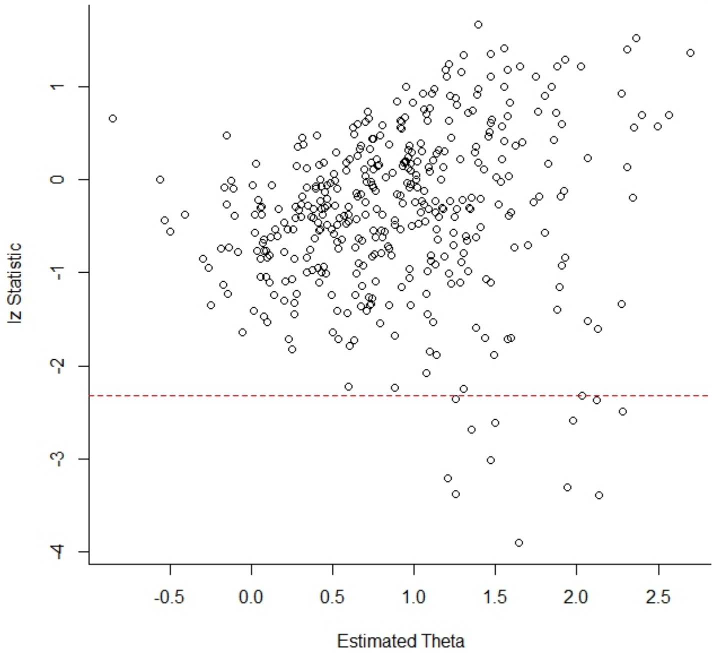


Figure B-9. Person fit distribution of the performance-based Geometry assessment, Form 014PO. Scores below the red line indicate poor fit.

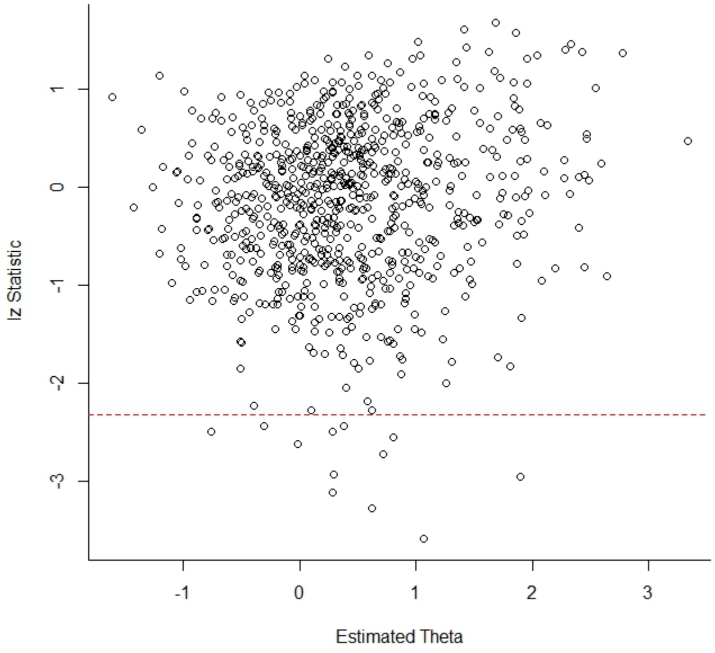


Figure B-10. Person fit distribution of the performance-based Integrated Math 1 assessment, Form 064PO. Scores below the red line indicate poor fit.

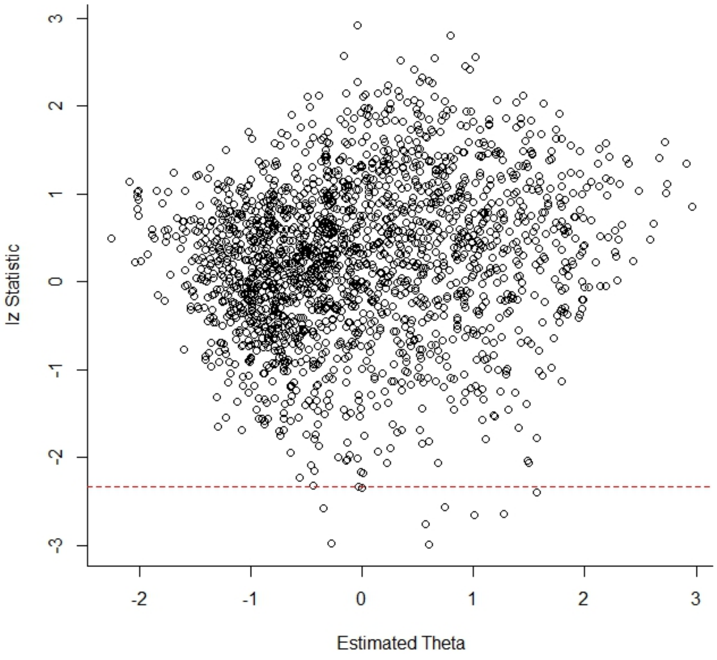


Figure B-11. Person fit distribution of the end-of-year ELA Grade 3 assessment, Form 124EO. Scores below the red line indicate poor fit.

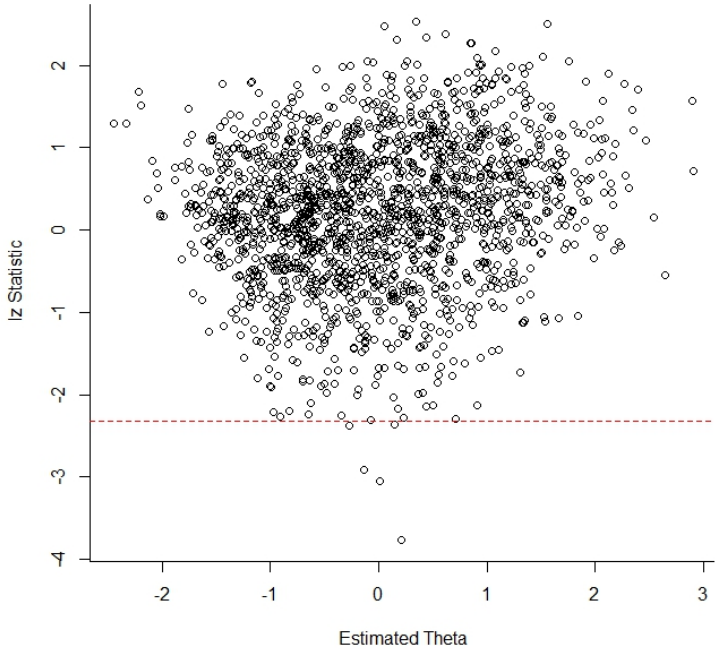


Figure B-12. Person fit distribution of the end-of-year ELA Grade 5 assessment, Form 064EP. Scores below the red line indicate poor fit.

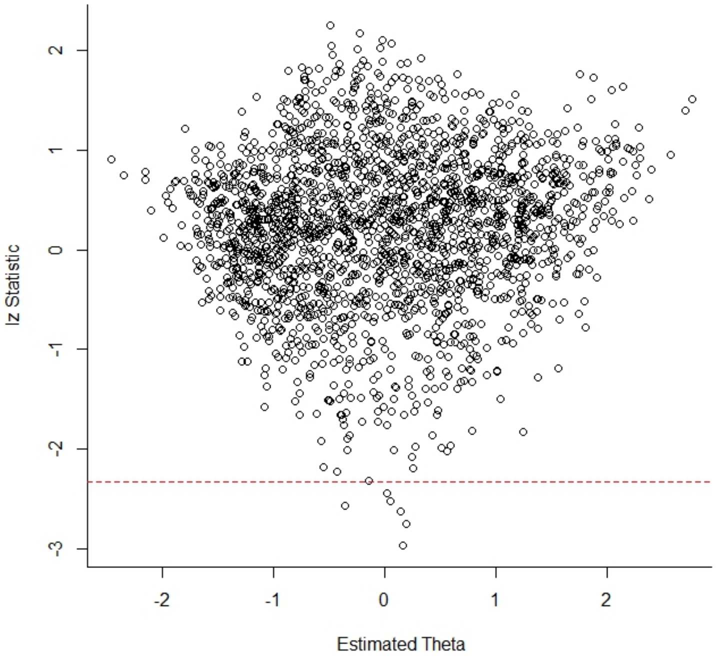


Figure B-13. Person fit distribution of the end-of-year ELA Grade 7 assessment, Form 064EP. Scores below the red line indicate poor fit.

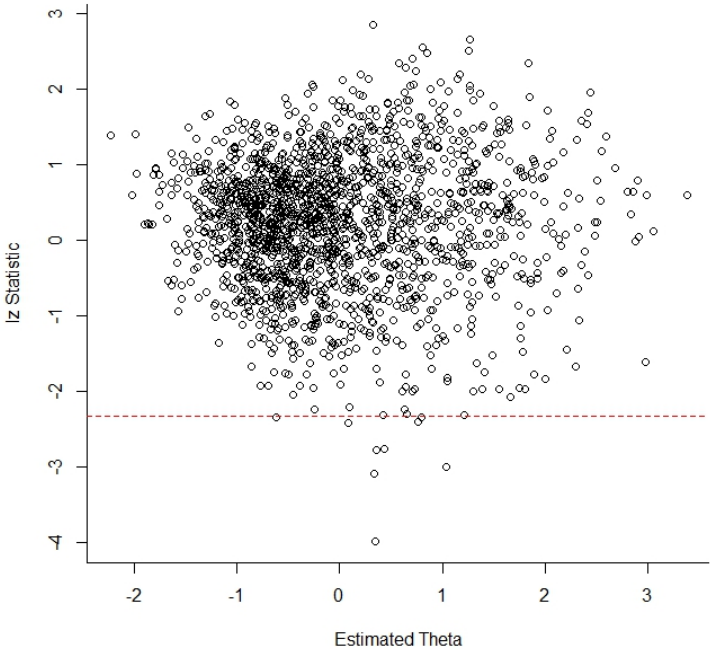


Figure B-14. Person fit distribution of the end-of-year ELA Grade 9 assessment, Form 064EP. Scores below the red line indicate poor fit.

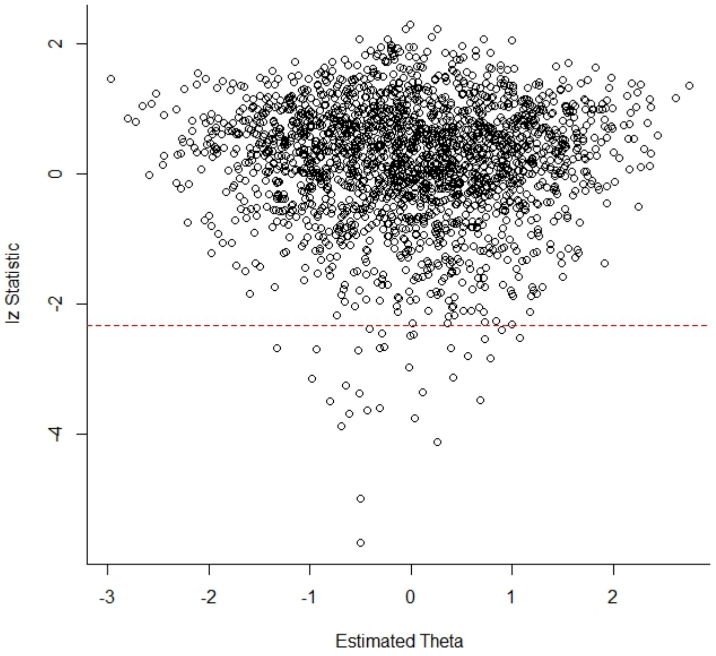


Figure B-15. Person fit distribution of the end-of-year Math Grade 3 assessment, Form 074EP. Scores below the red line indicate poor fit.

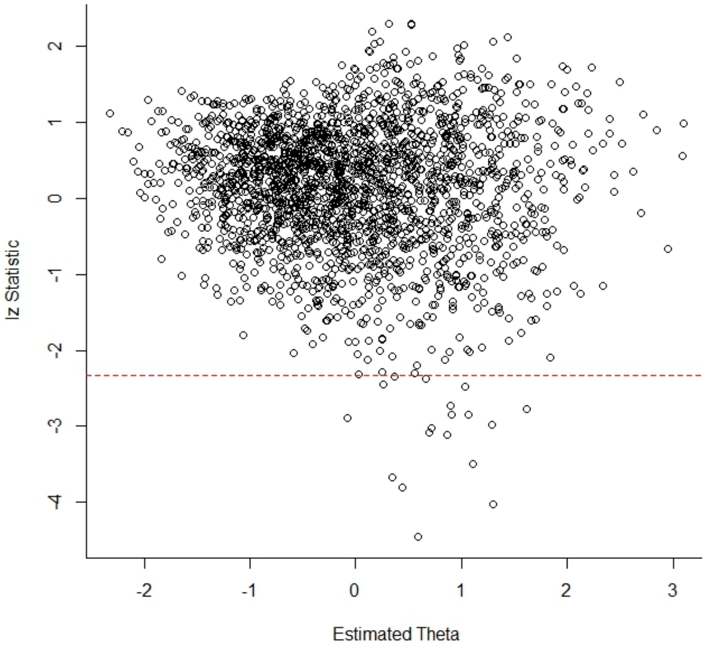


Figure B-16. Person fit distribution of the end-of-year Math Grade 5 assessment, Form 114EO. Scores below the red line indicate poor fit.

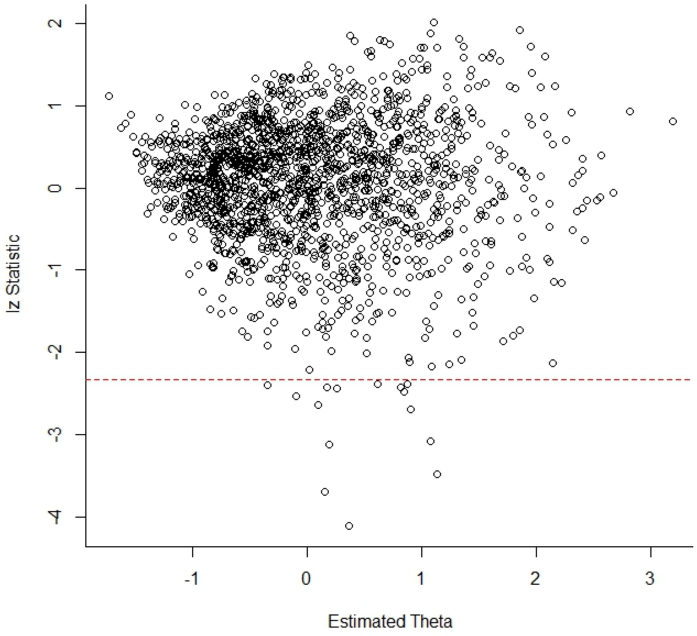


Figure B-17. Person fit distribution of the end-of-year Math Grade 7 assessment, Form 104EO. Scores below the red line indicate poor fit.

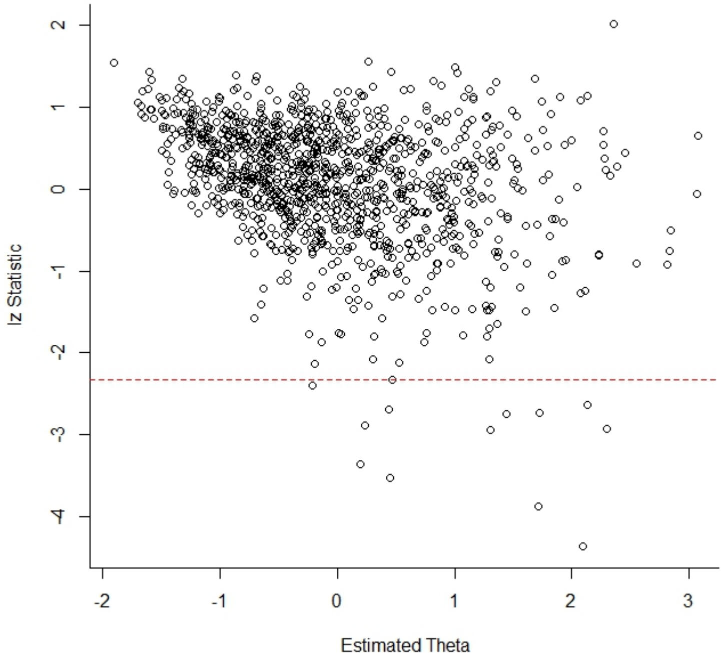


Figure B-18. Person fit distribution of the end-of-year Algebra 1 assessment, Form 024EO. Scores below the red line indicate poor fit.

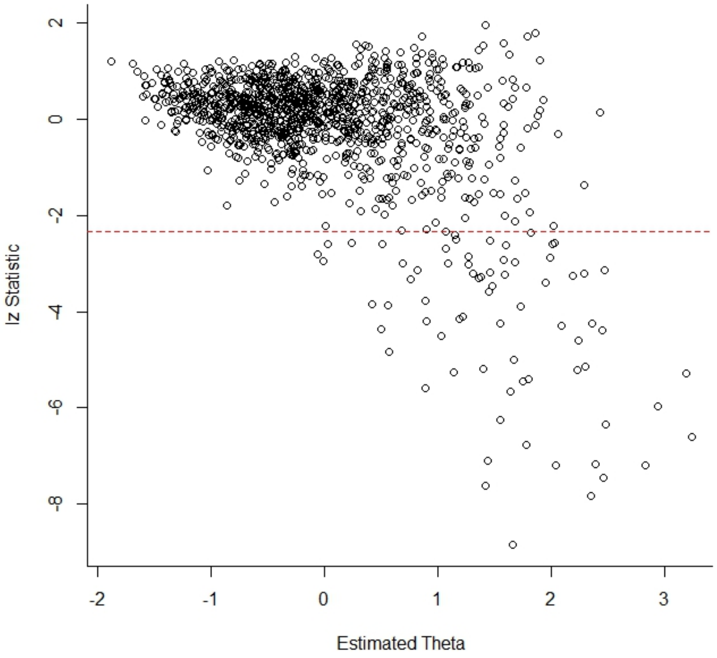


Figure B-19. Person fit distribution of the end-of-Year Math Geometry assessment, Form 114EO. Scores below the red line indicate poor fit.

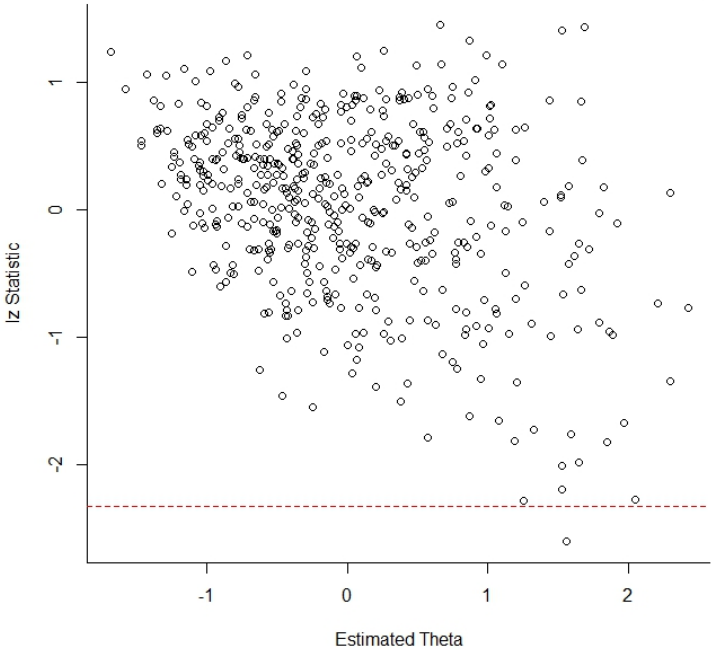


Figure B-20. Person fit distribution of the end-of-year Integrated Math 1 assessment, Form 024EO. Scores below the red line indicate poor fit.

Appendix C. Timing Analysis: Summary Tables

Table C-1. ELA EOY Timing Summary All Students (Grade 3 Only)*

	Typical Student's Time (Median)		Time for About 95% Completion (+ 2 SD)		Time for About 98% Completion (+ 2.5 SD)	
Grade 3	**No Extra Time**	***Extra Time***	**No Extra Time**	***Extra Time***	**No Extra Time**	***Extra Time***
Session 1 (N = 16/15) 70 mins, 105 mins	40.2	*42.5*	74.5	*85.7*	83.0	*96.5*
Session 2 (N = 16/15) 50 mins, 75 mins	29.0	*28.8*	56.5	*65.0*	63.4	*74.0*

*Forms omitted with fewer than 10 students.

Table C-2. ELA PBA Timing Summary All Students (Grade 4-5 Only)*

	Typical Student's Time (Median)		Time for About 95% Completion (+ 2 SD)		Time for About 98% Completion (+ 2.5 SD)	
Grade 4	**No Extra Time**	***Extra Time***	**No Extra Time**	***Extra Time***	**No Extra Time**	***Extra Time***
Literary Analysis (N = 21) 70 mins, 105 mins	38.7	*42.5*	75.1	*94.2*	84.2	*107.2*
Research Simulation (N = 21) 80 mins, 120 mins	42.8	*45.8*	83.3	*95.3*	93.5	*107.7*
Narrative Writing (N = 21) 50 mins, 75 mins	29.7	*29.7*	62.4	*69.6*	70.6	*79.6*
Grade 5						
Literary Analysis (N = 22/21) 70 mins, 105 mins	39.8	*45.0*	76.8	*90.2*	86.1	*101.5*
Research Simulation (N = 22/21) 70 mins, 105 mins	44.0	*48.6*	85.4	*101.0*	95.7	*114.2*
Narrative Writing (N = 22/21) 70 mins, 105 mins	29.0	*30.5*	61.4	*73.1*	69.5	*83.7*

*Forms omitted with fewer than 10 students.

Table C-3. ELA EOY Timing Summary All Students (Grade 4-5 Only)*

	Typical Student's Time (Median)		Time for About 95% Completion (+ 2 SD)		Time for About 98% Completion (+ 2.5 SD)	
Grade 4	**No Extra Time**	***Extra Time***	**No Extra Time**	**No Extra Time**	***Extra Time***	**No Extra Time**
Session 1 (N = 15) 70 mins, 105 mins	40.1	*44.9*	72.8	*85.4*	80.9	*95.5*
Session 2 (N = 15) 50 mins, 75 mins	29.0	*31.2*	54.9	*63.5*	61.3	*71.6*
Grade 5						
Session 1 (n = 17/15) 70 mins, 105 mins	39.7	*45.0*	72.7	*98.3*	81.0	*111.6*
Session 2 (n = 17/15) 70 mins, 105 mins	27.1	*27.4*	51.3	*65.5*	57.4	*75.1*

*Forms omitted with fewer than 10 students.

Table C-4. ELA PBA Timing Summary All Students (Grade 6-11 Only)*

	Typical Student's Time (Median)		Time for About 95% Completion (+ 2 SD)		Time for About 98% Completion (+ 2.5 SD)	
Grade 6	**No Extra Time**	***Extra Time***	**No Extra Time**	**No Extra Time**	***Extra Time***	**No Extra Time**
Literary Analysis (N = 21) 80 mins, 120 mins	40.7	*42.0*	79.2	*83.9*	88.8	*94.4*
Research Simulation (N = 21) 80 mins, 120 mins	43.9	*41.0*	85.7	*83.7*	96.2	*94.4*
Narrative Writing (N = 21) 50 mins, 75 mins	28.1	*26.7*	58.8	*57.8*	66.5	*65.5*
Grade 7						
Literary Analysis (N = 21) 80 mins, 120 mins	38.8	*41.8*	76.2	*85.1*	85.6	*95.9*
Research Simulation (N = 21) 80 mins, 120 mins	39.3	*40.2*	80.0	*84.0*	90.2	*95.0*
Narrative Writing (N = 21) 50 mins, 75 mins	24.2	*25.2*	52.2	*56.9*	59.2	*64.9*

(continued)

Table C-4. ELA PBA Timing Summary All Students (Grade 6-11 Only)* (continued)

	Typical Student's Time (Median)		Time for About 95% Completion (+ 2 SD)		Time for About 98% Completion (+ 2.5 SD)	
	No Extra Time	***Extra Time***	**No Extra Time**	**No Extra Time**	***Extra Time***	**No Extra Time**
Grade 8						
Literary Analysis (n=22) 80 mins, 120 mins	36.5	*41.3*	71.1	*81.7*	79.7	*91.8*
Research Simulation (n=22) 80 mins, 120 mins	40.1	*41.1*	79.9	*83.9*	89.8	*94.6*
Narrative Writing (n=22) 50 mins, 75 mins	22.8	*24.3*	49.1	*54.2*	55.7	*61.7*
Grade 9						
Literary Analysis (n=21) 80 mins, 120 mins	36.1	*37.8*	72.2	*75.2*	81.2	*84.6*
Research Simulation (n=21) 80 mins, 120 mins	35.9	*37.3*	76.5	*85.4*	86.7	*97.4*
Narrative Writing (n=21) 50 mins, 75 mins	22.4	*22.1*	50.5	*55.1*	57.5	*63.4*
Grade 10						
Literary Analysis (n=22/20) 80 mins, 120 mins	30.5	*31.6*	68.1	*68.4*	77.5	*77.6*
Research Simulation (n=22/20) 80 mins, 120 mins	33.0	*32.6*	74.8	*73.9*	85.2	*84.2*
Narrative Writing (n=22/20) 50 mins, 75 mins	19.2	*21.7*	49.4	*52.2*	57.0	*59.8*
Grade 11						
Literary Analysis (n=23/20) 80 mins, 120 mins	31.0	*31.2*	64.5	*69.1*	72.9	*78.5*
Research Simulation (n=23/20) 80 mins, 120 mins	29.9	*30.7*	67.0	*72.9*	76.3	*83.4*
Narrative Writing (n=23/20) 50 mins, 75 mins	18.5	*20.9*	43.5	*50.1*	49.7	*57.4*

*Forms omitted with fewer than 10 students.

Table C-5. ELA EOY Timing Summary All Students (Grade 6-11 Only)*

	Typical Student's Time (Median)		Time for About 95% Completion (+ 2 SD)		Time for About 98% Completion (+ 2.5 SD)	
Grade 6	**No Extra Time**	***Extra Time***	**No Extra Time**	**No Extra Time**	***Extra Time***	**No Extra Time**
Session 1 (n=15) 70 mins, 105 mins	34.0	*35.1*	61.7	*69.0*	68.6	*77.4*
Session 2 (n=15) 70 mins, 105 mins	33.8	*32.4*	64.0	*73.1*	71.6	*83.3*
Grade 7						
Session 1 (n=15) 70 mins, 105 mins	30.3	*34.1*	54.9	*63.2*	61.1	*70.4*
Session 2 (n=15) 70 mins, 105 mins	30.9	*31.8*	58.1	*62.9*	64.9	*70.7*
Grade 8						
Session 1 (n=15) 70 mins, 105 mins	29.1	*32.9*	54.0	*67.5*	60.3	*76.1*
Session 2 (n=15) 70 mins, 105 mins	30.0	*33.0*	57.2	*67.7*	64.0	*76.4*
Grade 9						
Session 1 (n=14) 70 mins, 105 mins	26.5	*26.2*	51.5	*56.6*	57.6	*64.2*
Session 2 (n=14) 70 mins, 105 mins	27.1	*26.0*	55.8	*61.5*	62.9	*70.4*
Grade 10						
Session 1 (n=14) 70 mins, 105 mins	25.6	*27.5*	51.9	*54.8*	58.4	*61.6*
Session 2 (n=14) 70 mins, 105 mins	25.8	*25.8*	56.0	*54.4*	63.6	*61.6*
Grade 11						
Session 1 (n=14) 70 mins, 105 mins	26.4	*27.2*	52.0	*61.9*	58.4	*70.6*
Session 2 (n=14) 70 mins, 105 mins	23.0	*22.2*	50.2	*54.3*	57.0	*62.3*

*Forms omitted with fewer than 10 students.

Table C-6. Math PBA Timing Summary (Grades 3-5)

	Typical Student's Time (Median)		Time for About 95% Completion (+ 2 SD)		Time for About 98% Completion (+ 2.5 SD)	
	No Extra Time	***Extra Time***	**No Extra Time**	***Extra Time***	**No Extra Time**	***Extra Time***
Grade 3						
Session 1 (N = 14/12) 50 mins, 75 mins	41.1	42.5	74.8	82.3	83.2	92.3
Session 2 (N = 14/12) 50 mins, 75 mins	33.4	33.9	63.2	67.5	70.6	75.9
Grade 4						
Session 1 (N = 15) 50 mins, 75 mins	43.0	44.6	74.0	87.7	81.7	98.5
Session 2 (N = 15) 50 mins, 75 mins	35.5	35.9	63.8	76.8	70.9	87.0
Grade 5						
Session 1 (n=12) 50 mins, 75 mins	48.7	53.0	81.6	103.7	89.8	116.4
Session 2 (n=12) 50 mins, 75 mins	35.8	37.3	65.5	87.3	72.9	99.8

*Forms omitted with fewer than 10 students.

Table C-7. Math EOY Timing Summary (Grades 3-5)

	Typical Student's Time (Median)		Time for About 95% Completion (+ 2 SD)		Time for About 98% Completion (+ 2.5 SD)	
	No Extra Time	***Extra Time***	**No Extra Time**	***Extra Time***	**No Extra Time**	***Extra Time***
Grade 3						
Session 1 (N = 12/12) 55 mins, 85 mins	37.4	40.0	69.2	76.1	77.1	85.1
Session 2 (N = 12/12) 55 mins, 85 mins	34.4	35.5	64.2	69.5	71.7	77.9
Grade 4						
Session 1 (N = 15/14) 55 mins, 85 mins	42.7	44.8	74.1	83.8	81.9	93.5
Session 2 (N = 15/14) 55 mins, 85 mins	39.4	42.2	70.0	76.8	77.7	85.4
Grade 5						
Session 1 (n=12) 55 mins, 85 mins	44.4	50.4	75.7	99.4	83.6	111.6
Session 2 (n=12) 55 mins, 85 mins	41.8	44.2	74.9	92.6	83.1	104.7

*Forms omitted with fewer than 10 students.

Table C-8. Math PBA Timing Summary (Grades 6-8)

	Typical Student's Time (Median)		Time for About 95% Completion (+ 2 SD)		Time for About 98% Completion (+ 2.5 SD)	
	No Extra Time	***Extra Time***	**No Extra Time**	***Extra Time***	**No Extra Time**	***Extra Time***
Grade 6						
Session 1 (n=16/15) 50 mins, 75 mins	46.3	49.4	79.9	93.9	88.3	105.0
Session 2 (n=16/15) 50 mins, 75 mins	34.1	34.7	63.0	70.6	70.3	79.6
Grade 7						
Session 1 (n=18/11) 50 mins, 75 mins	49.0	49.7	86.8	92.1	96.2	102.6
Session 2 (n=18/11) 50 mins, 75 mins	37.4	39.1	69.5	80.3	77.5	90.6
Grade 8						
Session 1 (n=15/14) 50 mins, 75 mins	41.2	43.5	71.8	82.5	79.4	92.3
Session 2 (n=15/14) 50 mins, 75 mins	31.1	32.1	61.6	67.1	69.2	75.8

*Forms omitted with fewer than 10 students.

Table C-9. Math EOY Timing Summary (Grades 6-8)

	Typical Student's Time (Median)		Time for About 95% Completion (+ 2 SD)		Time for About 98% Completion (+ 2.5 SD)	
	No Extra Time	***Extra Time***	**No Extra Time**	***Extra Time***	**No Extra Time**	***Extra Time***
Grade 6						
Session 1 (n=14) 60 mins, 90 mins	50.5	51.5	85.1	96.2	93.7	107.4
Session 2 (n=14) 60 mins, 90 mins	33.7	32.4	63.9	69.7	71.4	79.1
Grade 7						
Session 1 (n=17/14) 60 mins, 90 mins	41.3	41.8	72.9	80.4	80.9	90.1
Session 2 (n=17/14) 60 mins, 90 mins	30.0	29.0	58.4	61.0	65.5	69.0
Grade 8						
Session 1 (n=14) 60 mins, 90 mins	37.4	39.6	69.1	72.3	77.0	80.4
Session 2 (n=14) 60 mins, 90 mins	30.6	29.9	62.8	64.6	70.9	73.3

*Forms omitted with fewer than 10 students.

Table C-10. Math PBA Timing Summary (High School)

	Typical Student's Time (Median)		Time for About 95% Completion (+ 2 SD)		Time for About 98% Completion (+ 2.5 SD)	
	No Extra Time	*Extra Time*	No Extra Time	*Extra Time*	No Extra Time	*Extra Time*
Algebra 1						
Session 1 (N = 17/4) 55 mins, 85 mins	52.5	57.2	86.7	101.0	95.2	112.0
Session 2 (N = 17/4) 55 mins, 85 mins	28.4	28.7	57.8	61.2	65.1	69.3
Geometry						
Session 1 (N = 16/3) 55 mins, 85 mins	43.9	49.9	75.2	87.9	83.0	97.4
Session 2 (N = 16/3) 55 mins, 85 mins	30.5	31.2	59.4	61.2	66.7	68.7
Algebra 2						
Session 1 (N = 14/1) 65 mins, 100 mins	50.3	53.1	84.7	115.7	93.3	131.4
Session 2 (N = 14/1) 65 mins, 100 mins	32.4	27.2	64.4	59.3	72.4	67.3
Math I						
Session 1 (N = 2/1) 55 mins, 85 mins	46.7	50.8	83.3	95.9	92.4	107.2
Session 2 (N = 2/1) 55 mins, 85 mins	29.1	41.3	59.7	106.0	67.4	122.2
Math II						
Session 1 (N = 2) 60 mins, 90 mins	44.4	NA	76.2	NA	84.2	NA
Session 2 (N = 2) 60 mins, 90 mins	28.0	NA	53.6	NA	60.0	NA
Math III						
Session 1 (N = 2/1) 65 mins, 100 mins	61.9	60.6	109.7	96.2	121.7	105.1
Session 2 (N = 2/1) 65 mins, 100 mins	34.4	37.9	66.4	67.6	74.5	75.1

*Forms omitted with fewer than 10 students.

Table C-11. Math EOY Timing Summary (High School)

	Typical Student's Time (Median)		Time for About 95% Completion (+ 2 SD)		Time for About 98% Completion (+ 2.5 SD)	
	No Extra Time	***Extra Time***	**No Extra Time**	***Extra Time***	**No Extra Time**	***Extra Time***
Algebra 1						
Session 1 (N = 16/13) 70 mins, 105 mins	39.0	38.6	76.1	80.9	85.3	91.4
Session 2 (N = 16/13) 70 mins, 105 mins	40.9	40.5	54.4	58.1	62.0	66.6
Geometry						
Session 1 (N = 15/12) 70 mins, 105 mins	38.8	41.0	73.4	89.9	82.1	102.1
Session 2 (N = 15/12) 70 mins, 105 mins	22.3	26.3	52.7	64.0	60.4	73.4
Algebra 2						
Session 1 (N = 13/12) 70 mins, 105 mins	38.6	36.7	80.0	82.7	90.3	94.2
Session 2 (N = 13/12) 70 mins, 105 mins	19.6	20.0	50.1	54.7	57.7	63.4
Math I						
Session 1 (N = 3/2) 70 mins, 105 mins	41.3	36.7	83.3	70.9	93.8	79.4
Session 2 (N = 3/2) 70 mins, 105 mins	22.4	15.7	53.6	36.7	61.4	42.0
Math II						
Session 1 (N = 3) 70 mins, 105 mins	35.9	NA	71.2	NA	80.0	NA
Session 2 (N = 3)	18.5	NA	42.8	NA	48.9	NA
Math III						
Session 1 (N = 3/2) 70 mins, 105 mins	38.8	36.6	84.5	71.6	95.9	80.3
Session 2 (N = 3/2) 70 mins, 105 mins	17.4	15.1	46.1	37.4	53.3	42.9

*Forms omitted with fewer than 10 students.

Appendix D. Timing Analysis: Percentile Data

Table D-1. Percentile Data for the PBA and EOY Assessments by Subject and Grade.

PBA																	
ELA																	
	Obs	**Grade**	**QR1**	**QR2**	**QR3**	**session1_25**	**session1_75**	**session1_95**	**session1_98**	**session2_25**	**session2_75**	**session2_95**	**session2_98**	**session3_25**	**session3_75**	**session3_95**	**session3_98**
	1	3	23.2333	25.0333	17.1333	26.6	49.8333	73.05	85.75	31.95	56.9833	79.7833	93.383	9.9167	27.05	44.5667	51.65
	2	4	23.55	28.0167	23.3833	28.8833	52.4333	73.5833	88.55	31.2333	59.25	82.6667	95.85	19.6333	43.0167	61.7	72.65
	3	5	23.8167	27.7833	21.9833	29.7167	53.5333	77.3833	92.7167	32.5833	60.3667	86.1833	102.1	19.7	41.6833	61.9167	73.05
	4	6	25.1	27.7	21.0667	30.3833	55.4833	80.55	92.1833	32.1833	59.8833	85.1	99.95	19.2333	40.3	59.3167	70.4
	5	7	23.8833	25.75	17.9833	28.8333	52.7167	78.7667	89.7	28.6	54.35	80.05	93.217	16.6	34.5833	51.9167	61.15
	6	8	21.6167	25.9167	16.8167	27.2167	48.8333	72.3667	82.75	28.6833	54.6	79.4667	92.4	15.3667	32.1833	50.15	59.1833
	7	9	22.65	26.15	17.2333	25.8833	48.5333	73.4833	85.2667	24.4333	50.5833	77.4167	90.617	14.5833	31.8167	50.6667	60.4
	8	10	23.675	27.1667	18.95	20.1167	43.7917	70.0667	81.3833	21.2	48.3667	74.8167	85.6	10.9333	29.8833	49.6167	59.6667
	9	11	21.8167	25.4667	16.7167	20.7167	42.5333	66.6833	78.1	19.05	44.5167	71.45	81.583	11.15	27.8667	45.9833	53.6667
ALG I																	
	Obs	**QR1**	**QR2**	**session 1_25**	**session1_75**	**session1_95**	**session1_98**	**session2_25**	**session2_75**	**session2_95**	**session2_98**						
	1	23.4333	19.4333	40.2667	63.7	83.1833	90.75	19.5333	38.9667	58.5667	67.4167						
ALG II																	
	Obs	**QR1**	**QR2**	**session 1_25**	**session1_75**	**session1_95**	**session1_98**	**session2_25**	**session2_75**	**session2_95**	**session2_98**						
	1	22.9417	22.05	38.6583	61.6	82.8167	93.8167	21.5333	43.5833	64.1667	73.6167						
GEO																	
	Obs	**QR1**	**QR2**	**session 1_25**	**session1_75**	**session1_95**	**session1_98**	**session2_25**	**session2_75**	**session2_95**	**session2_98**						
	1	20.7	20.15	34.1333	54.8333	74.8	82.4667	20.2667	40.4167	58.5333	68.55						
MATH 3-8																	
	Obs	**Grade**	**QR1**	**QR2**	**session1_25**	**session1_75**	**session1_95**	**session1_98**	**session2_25**	**session2_75**	**session2_95**	**session2_98**					
	1	3	21.5333	19.7167	31.7	53.2333	74.1333	83.05	24.9833	44.7	65.8	75.4167					
	2	4	21.45	19.0667	33.3333	54.7833	74.05	81.8667	26.35	45.4167	64.1333	71.9667					
	3	5	23.1333	19.5667	38.8	61.9333	79.35	90.7667	27.1333	46.7	65.7	73.3167					
	4	6	21.6333	19.7667	37.1667	58.8	77	89.85	25.7333	45.5	63.9833	72.45					
	5	7	21.6	20.65	40.7167	62.3167	79.5333	92.0333	28.1333	48.7833	66.95	73.85					
	6	8	18.9333	19.6333	33.2333	52.1667	70.85	78.7333	22.9833	42.6167	60.3667	68.8333					
MATH I																	
	Obs	**QR1**	**QR2**	**session 1_25**	**session1_75**	**session1_95**	**session1_98**	**session2_25**	**session2_75**	**session2_95**	**session2_98**						
	1	21.8833	21.7667	36.2833	58.1667	84.7833	100.333	22.7167	44.4833	71.1	82.55						
MATH II																	
	Obs	**QR1**	**QR2**	**session 1_25**	**session1_75**	**session1_95**	**session1_98**	**session2_25**	**session2_75**	**session2_95**	**session2_98**						
	1	21.8333	18.3	36.7	58.5333	77.9167	85.6167	21.9667	40.2667	56.4833	63.05						
MATH III																	
	Obs	**QR1**	**QR2**	**session 1_25**	**session1_75**	**session1_95**	**session1_98**	**session2_25**	**session2_75**	**session2_95**	**session2_98**						
	1	25.6	18.3833	46.8333	72.4333	95.5833	116.317	25.2167	43.6	63.85	70.1167						

EOY

ELA

Obs	Grade	QR1	QR2	session1_25	session1_75	session1_95	session1_98	session2_25	session2_75	session2_95	session2_98
1	3	21.3333	17.0333	30.9167	52.25	73.6667	86.4667	21.2	38.2333	54.7833	65
2	4	20.5333	16.7	31.4167	51.95	72.5333	83.7167	21.55	38.25	53.5333	61.8167
3	5	20.3833	14.9833	30.75	51.1333	71.5833	85.15	20.1667	35.15	50.1833	59.9333
4	6	17.8667	19.1667	26.15	44.0167	63.0833	71.1667	24.7333	43.9	63.8833	72.05
5	7	15.3833	17.1167	23.3833	38.7667	55.3333	63.9833	22.45	39.5667	56.6167	65.1
6	8	14.0917	17.6333	21.4083	35.5	51.1	60.1333	19.85	37.4833	53.9167	62.45
7	9	15.35	19.4333	19.3167	34.6667	51.1333	60.4167	16.8833	36.3167	54.3667	64.25
8	10	15.4833	20.2167	18.4333	33.9167	50.55	60.5	15.7167	35.9333	56	65
9	11	16.8	20.1833	17.7833	34.5833	51.85	60.85	12.1167	32.3	49.25	57.2833
10	12	18.6917	23.0083	17.7583	36.45	59.5333	66.7	8.3333	31.3417	53.0667	64.8

ALG I

Obs	QR1	QR2	session1_25	session1_75	session1_95	session1_98	session2_25	session2_75	session2_95	session2_98
1	24.4833	20.3167	26.9667	51.45	73.8333	86.15	13.15	33.4667	53.3	63.55

ALG II

Obs	QR1	QR2	session1_25	session1_75	session1_95	session1_98	session2_25	session2_75	session2_95	session2_98
1	29.4167	21.3667	24.0167	53.4333	76.6167	88.35	9.66667	31.0333	50.2333	60.0833

GEO

Obs	QR1	QR2	session1_25	session1_75	session1_95	session1_98	session2_25	session2_75	session2_95	session2_98
1	22.8167	22.0333	28.05	50.8667	70.4	82.4833	12.2	34.2333	53.15	61.7333

MATH 3-8

Obs	Grade	QR1	QR2	session1_25	session1_75	session1_95	session1_98	session2_25	session2_75	session2_95	session2_98
1	3	21.5333	19.7167	31.7	53.2333	74.1333	83.05	24.9833	44.7	65.8	75.4167
2	4	21.45	19.0667	33.3333	54.7833	74.05	81.8667	26.35	45.4167	64.1333	71.9667
3	5	23.1333	19.5667	38.8	61.9333	79.35	90.7667	27.1333	46.7	65.7	73.3167
4	6	21.6333	19.7667	37.1667	58.8	77	89.85	25.7333	45.5	63.9833	72.45
5	7	21.6	20.65	40.7167	62.3167	79.5333	92.0333	28.1333	48.7833	66.95	73.85
6	8	18.9333	19.6333	33.2333	52.1667	70.85	78.7333	22.9833	42.6167	60.3667	68.8333

MATH I

Obs	QR1	QR2	session1_25	session1_75	session1_95	session1_98	session2_25	session2_75	session2_95	session2_98
1	29.825	20.5667	27.1583	56.9833	87.8333	97.0667	12.6083	33.175	56.2333	66.65

MATH II

Obs	QR1	QR2	session1_25	session1_75	session1_95	session1_98	session2_25	session2_75	session2_95	session2_98
1	30.0167	17.6833	17.6833	47.7	71.2333	84.0667	7.78333	25.4667	41.4667	49.9333

MATH III

Obs	QR1	QR2	session1_25	session1_75	session1_95	session1_98	session2_25	session2_75	session2_95	session2_98
1	30.5167	19.3	24.3667	54.8833	88.2667	104	8.85	28.15	48.8333	56.55

Appendix E. Administrator Survey Results

PBA ELA Administrator Survey Results

1. Indicate the proportion of students who needed the Additional Time Allowed.
   - None *(Value =1)*
   - Less than a third of students *(Value =2)*
   - Approximately half of students *(Value =3)*
   - More than half of students *(Value =4)*

1a. Please indicate how much additional time was used by the majority of these students.

- Less than 10 minutes *(Value =1)*
- Between 10-20 minutes *(Value =2)*
- More than 20 minutes *(Value =3)*

Table E-1. PBA ELA Administrator Survey Results By Grade

Grade	Proportion of Students Who Needed Additional Time				Amount of Additional Time Used		
	None	Less than a Third	Approx. Half	More Than Half	Less Than 10 Mins	10-20 Mins	More Than 20 Mins
3	105 (23.8%)	241 54.6%	60 (13.6%)	35 (7.9%)	86 (25.7%)	188 (56.1%)	61 (18.2%)
4	125 (31.7%)	197 (50.0%)	38 (9.6%)	34 (8.6%)	74 (27.8%)	140 (52.6%)	52 (19.5%)
5	90 (28.8%)	155 (49.5)	44 (14.1%)	24 (7.7%)	41 (18.8%)	136 (62.4%)	41 (18.8%)
6	94 (34.2%)	149 (54.2%)	16 (5.8%)	16 (5.8%)	58 (32.6%)	83 (46.6%)	37 (20.8%)
7	149 (47.5%)	134 (42.7%)	16 (5.1%)	15 (4.8%)	54 (33.3%)	81 (50.0%)	27 (16.7%)
8	97 (42.0%)	99 (42.9%)	18 (7.8%)	17 (7.4%)	33 (24.8%)	75 (56.4%)	25 (18.8%)
9	117 (55.5%)	75 (35.5%)	9 (4.3%)	10 (4.7%)	35 (37.6%)	44 (47.3%)	14 (15.1%)
10	83 (55.0%)	57 (37.7%)	5 (3.3%)	6 (4.0%)	21 (31.3%)	32 (47.8%)	14 (20.9%)
11	133 (60.7%)	72 (32.9%)	4 (1.8%)	10 (4.6%)	35 (41.2%)	35 (41.2%)	15 (17.6%)

Note. Percentages are based on valid percent. Individuals with missing data were excluded from analyses.

EOY ELA Administrator Survey Results

2. Indicate the proportion of students who needed the Additional Time Allowed.
   - None *(Value =1)*
   - Less than a third of students *(Value =2)*

- Approximately half of students *(Value =3)*
- More than half of students *(Value =4)*

1a. Please indicate how much additional time was used by the majority of these students.

- Less than 10 minutes *(Value =1)*
- Between 10-20 minutes *(Value =2)*
- More than 20 minutes *(Value =3)*

Table E-2. EOY ELA Administrator Survey Results By Grade

Grade	Proportion of Students Who Needed Additional Time				Amount of Additional Time Used		
	None	Less than a Third	Approx. Half	More Than Half	Less Than 10 Mins	10-20 Mins	More Than 20 Mins
3	35 (23.6%)	82 (55.4%)	19 (12.8%)	12 (8.1%)	22 (19.6%)	64 (57.1%)	26 (23.2%)
4	35 (24.6%)	79 (80.3%)	21 (14.8%)	7 (4.9%)	29 (26.9%)	56 (51.9%)	23 (21.3%)
5	38 (28.1%)	63 (46.7%)	20 (14.8%)	14 (10.4%)	20 (21.3%)	52 (55.3%)	22 (23.4%)
6	59 (46.8%)	56 (44.4%)	6 (4.8%)	5 (4.0%)	22 (33.3%)	35 (53.0%)	9 (13.6%)
7	63 (52.5%)	47 (39.2%)	4 (3.3%)	6 (5.0%)	20 (35.1%)	25 (43.9%)	12 (21.1%)
8	41 (50.6%)	33 (40.7%)	4 (4.9%)	3 (3.7%)	11 (27.5%)	22 (55.0%)	7 (17.5%)
9	68 (68.0%)	29 (29.0%)	0 (0.0%)	3 (3.0%)	14 (42.4%)	14 (42.4%)	5 (15.2%)
10	71 (69.6%)	24 (23.5%)	3 (2.9%)	4 (3.9%)	12 (37.5%)	14 (43.8%)	6 (18.8%)
11	50 (71.4%)	17 (24.3%)	1 (1.4%)	2 (2.9%)	10 (47.6%)	8 (38.1%)	3 (14.3%)

Note. Percentages are based on valid percent. Individuals with missing data were excluded from analyses.

PBA Math Administrator Survey Results

3. Indicate the proportion of students who needed the Additional Time Allowed.
   - None *(Value =1)*
   - Less than a third of students *(Value =2)*
   - Approximately half of students *(Value =3)*
   - More than half of students *(Value =4)*

1a. Please indicate how much additional time was used by the majority of these students.

- Less than 10 minutes *(Value =1)*
- Between 10-20 minutes *(Value =2)*
- More than 20 minutes *(Value =3)*

Table E-3. PBA Math Administrator Survey Results By Grade

Grade	Proportion of Students Who Needed Additional Time				Amount of Additional Time Used		
	None	Less than a Third	Approx. Half	More Than Half	Less Than 10 Mins	10-20 Mins	More Than 20 Mins
3	31 (11.6%)	146 (66.3%)	62 (23.2%)	28 (10.5%)	37 (15.5%)	148 (62.2%)	53 (22.3%)
4	45 (13.1%)	170 (49.6%)	65 (19.0%)	63 (18.4%)	41 (13.9%)	166 (56.1%)	89 (30.1%)
5	19 (10.7%)	89 (50.3%)	31 (17.5%)	38 (21.5%)	26 (16.8%)	92 (59.4%)	37 (23.9%)
6	36 (15.4%)	128 (54.7%)	38 (16.2%)	32 (13.7%)	33 (16.9%)	106 (54.4%)	56 (28.7%)
7	30 (13.6%)	113 (51.4%)	46 (20.9%)	31 (14.1%)	38 (19.8%)	113 (58.9%)	41 (21.4%)
8	45 (26.9%)	81 (48.5%)	21 (12.6%)	20 (12.0%)	31 (25.6%)	66 (54.5%)	24 (19.8%)
Algebra 1	38 (32.2%)	52 (44.1%)	15 (12.7%)	13 (11.0%)	23 (29.1%)	42 (53.2%)	14 (17.7%)
Algebra 2	52 (34.0%)	76 (49.7%)	15 (9.8%)	10 (6.5%)	33 (32.7%)	43 (42.6%)	25 (24.8%)
Geometry	65 (38.9%)	75 (44.9%)	18 (10.8%)	9 (5.4%)	42 (42.4%)	35 (35.4%)	22 (22.2%
Int. Math 1	12 (35.3%)	20 (58.8%)	1 (2.9%)	1 (2.9%)	6 (26.1%)	12 (52.2%)	5 (21.7%)
Int. Math 2	5 (41.7%)	7 (58.3%)	0 (0.0%)	0 (0.0%)	3 (42.9%)	3 (42.9%)	1 (14.3%)
Int. Math 3	7 (23.3%)	21 (70.0%)	2 (6.7%)	0 (0.0%)	10 (41.7%)	12 (50.0%)	2 (8.3%)

Note. Percentages are based on valid percent. Individuals with missing data were excluded from analyses.

EOY Math Administrator Survey Results

4. Indicate the proportion of students who needed the Additional Time Allowed.
   - ○ None *(Value =1)*
   - ○ Less than a third of students *(Value =2)*
   - ○ Approximately half of students *(Value =3)*
   - ○ More than half of students *(Value =4)*

1a. Please indicate how much additional time was used by the majority of these students.

- ○ Less than 10 minutes *(Value =1)*
- ○ Between 10-20 minutes *(Value =2)*
- ○ More than 20 minutes *(Value =3)*

Table E-3. EOY Math Administrator Survey Results By Grade

Grade	Proportion of Students Who Needed Additional Time				Amount of Additional Time Used		
	None	Less than a Third	Approx. Half	More Than Half	Less Than 10 Mins	10-20 Mins	More Than 20 Mins
3	22 (14.0%)	94 (59.9%)	28 (17.8%)	13 (8.3%)	21 (15.7%)	82 (61.2%)	31 (23.1%)
4	23 (14.3%)	92 (57.1%)	32 (19.9%)	14 (8.7%)	23 (16.9%)	85 (62.5%)	28 (20.6%)
5	17 (13.8%)	67 (54.5%)	20 (16.3%)	19 (15.4%)	15 (14.4%)	60 (57.7%)	29 (27.9%)
6	11 (10.7%)	62 (60.2%)	15 (14.6%)	15 (14.6%)	15 (16.5%)	47 (51.6%)	29 (31.9%)
7	23 (19.2%)	71 (59.2%)	14 (11.7%)	12 (10.0%)	20 (20.6%)	56 (57.7%)	21 (21.6%)
8	32 (36.0%)	41 (46.1%)	8 (9.0%)	8 (9.0%)	14 (24.6%)	31 (54.4%)	12 (21.1%)
Algebra 1	35 (31.8%)	62 (56.4%)	6 (5.5%)	7 (6.4%)	20 (27.0%)	40 (54.1%)	14 (18.9%)
Algebra 2	35 (44.3%)	36 (45.6%)	5 (6.3%)	3 (3.8%)	12 (27.3%)	26 (59.1%)	6 (13.6%)
Geometry	36 (52.9%)	28 (41.2%)	3 (4.4%)	1 (1.5%)	14 (42.4%)	12 (36.4%)	7 (21.2%)
Int. Math 1	4 (36.4%)	4 (36.4%)	1 (9.1%)	2 (18.2%)	2 (25.0%)	2 (25.0%)	4 (50.0%)
Int. Math 2	1 (50.0%)	1 (50.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (100.0%)	0 (0.0%)
Int. Math 3	7 (23.3%)	21 (70.0%)	2 (6.7%)	0 (0.0%)	10 (41.7%)	12 (50.0%)	2 (8.3%)

Note. Percentages are based on valid percent. Individuals with missing data were excluded from analyses.

Appendix F. Student Survey Results

ELA PBA Student Survey Results

1. Did you have enough time to finish this test? **Variable Name = CBT_ELA_Time_Finish**

I finished very early	**Value = 1**
I finished on time	**Value = 2**
I had to rush to finish	**Value = 3**
I did not finish	**Value = 4**

Table F-1. PBA ELA Student Survey Results by Grade

Grade	Did you have enough time to finish this test?			
	Early	On Time	Rush to Finish	Did Not Finish
3	10,680 38.5%	14,958 53.9%	1,240 4.5%	874 3.1%
4	10,982 38.2%	16,133 56.1%	1,070 3.7%	586 2.0%
5	9,657 37.0%	15,100 57.8%	948 3.6%	427 1.6%
6	10,045 37.9%	15,172 57.3%	976 3.7%	308 1.2%
7	11,331 42.6%	13,844 52.0%	1,167 4.4%	261 1.0%
8	10,981 43.7%	12,764 50.8%	1,130 4.5%	237 0.9%
9	9,218 40.1%	11,888 51.7%	1,538 6.7%	359 1.6%
10	7,449 39.4%	9,521 50.4%	1,569 8.3%	353 1.9%
11	8,551 48.1%	7,724 43.4%	1,228 6.9%	285 1.6%

Note. Percentages are based on valid percent. Individuals with missing data were excluded from analyses.

ELA EOY Student Survey Results

2. Did you have enough time to finish this test? **Variable Name = CBT_ELA_Time_Finish**

I finished very early **Value = 1**

I finished on time **Value = 2**

I had to rush to finish **Value = 3**

I did not finish **Value = 4**

Table F-2. EOY ELA Student Survey Results by Grade

Grade	Did you have enough time to finish this test?			
	Early	On Time	Rush to Finish	Did Not Finish
3	9,109 40.7%	11,846 53.0%	922 4.1%	495 2.2%
4	8,842 39.6%	12,275 55.0%	878 3.9%	338 1.5%
5	7,781 38.8%	11,321 56.4%	766 3.8%	212 1.1%
6	8,484 40.9%	11,154 53.8%	918 4.4%	178 0.9%
7	8,337 42.7%	10,020 51.3%	995 5.1%	163 0.8%
8	7,303 45.8%	7,598 47.7%	895 5.6%	148 0.9%
9	6,405 41.1%	7,644 49.0%	1,337 8.6%	204 1.3%
10	5,247 41.0%	6,098 47.7%	1,265 9.9%	179 1.4%
11	4,913 44.2%	5,030 45.3%	988 8.9%	173 1.6%

Note. Percentages are based on valid percent. Individuals with missing data were excluded from analyses.

Math PBA Student Survey Results

5. Did you have enough time to finish this test? **Variable Name = CBT_Math_Time_Finish**

I finished very early **Value = 1**

I finished on time **Value = 2**

I had to rush to finish **Value = 3**

I did not finish **Value = 4**

Table F-3. PBA Math Student Survey Results by Grade

Grade	Did you have enough time to finish this test?			
	Early	On Time	Rush to Finish	Did Not Finish
3	7,661 34.8%	11,661 52.9%	1,459 6.6%	1,244 5.6%
4	6,786 30.0%	13,055 57.7%	1,477 6.5%	1,301 5.8%
5	5,714 27.6%	12,386 59.9%	1,479 7.2%	1,100 5.3%
6	5,491 25.1%	13,488 61.5%	1,972 9.0%	969 4.4%
7	4,729 22.2%	12,921 60.6%	2,535 11.9%	1,151 5.4%
8	4,930 26.2%	11,590 61.7%	1,745 9.3%	529 2.8%
Algebra 1	4,399 26.5%	9,585 57.8%	1,852 11.2%	758 4.6%
Algebra 2	5,249 32.0%	8,069 49.1%	2,158 13.1%	951 5.8%
Geometry	5,656 32.8%	9,032 52.3%	1,929 11.2%	652 3.8%
Int. Math 1	580 29.8%	1,024 52.5%	249 12.8%	96 4.9%
Int. Math 2	523 34.4%	750 49.4%	190 12.5%	56 3.7%
Int. Math 3	403 29.1%	722 52.1%	185 13.3%	76 5.5%

Note. Percentages are based on valid percent. Individuals with missing data were excluded from analyses.

Math EOY Student Survey Results

6. Did you have enough time to finish this test? **Variable Name = CBT_Math_Time_Finish**

I finished very early **Value = 1**

I finished on time **Value = 2**

I had to rush to finish **Value = 3**

I did not finish **Value = 4**

Table F-4. EOY Math Student Survey Results by Grade

Grade	Did you have enough time to finish this test?			
	Early	On Time	Rush to Finish	Did Not Finish
3	8,894 40.5%	11,535 52.5%	996 4.5%	537 2.4%
4	7,232 34.0%	12,330 58.0%	1,153 5.4%	561 2.6%
5	5,354 29.2%	11,248 61.3%	1,207 6.6%	541 2.9%
6	5,383 28.9%	11,117 59.7%	1,694 9.1%	424 2.3%
7	6,013 31.3%	11,397 59.3%	1,512 7.9%	298 1.6%
8	4,763 30.8%	8,992 58.1%	1,463 9.5%	252 1.6%
Algebra 1	4,668 33.3%	7,605 54.2%	1,482 10.6%	277 2.0%
Algebra 2	4,385 35.5%	5,821 47.4%	1,672 13.6%	436 3.5%
Geometry	4,206 33.5%	6,468 51.4%	1,564 12.4%	334 2.7%
Int. Math 1	493 30.9%	874 54.7%	189 11.8%	42 2.6%
Int. Math 2	376 36.9%	487 47.7%	130 12.7%	27 2.6%
Int. Math 3	278 30.3%	488 53.3%	128 14.0%	22 2.4%

Note. Percentages are based on valid percent. Individuals with missing data were excluded from analyses.

Appendix G. Split Half Reliability Estimates

Table G-1. Split Half Data for Number Correct (NC) and Log Correct Response Time (LCT) – All Students

		Number of Cases										
		All	NC > 0		NC Means		LCT Means		Odd-Even Corr.		Reliability Estimate	
Grade	Cond.	Students	Odds	Evens	Odds	Evens	Odds	Evens	NC	LCT	NC	LCT
3	End	1,290	1,059	936	1.94	1.53	3.85	3.70	0.93	0.25	0.96	0.40
4 Odds	End-Timed	1,339	1,131	985	2.00	1.56	3.84	3.64	0.92	0.20	0.96	0.33
4 Evens	Interspersed	1,924	1,757	1,539	2.10	1.63	4.15	4.05	0.91	0.31	0.95	0.47
4	End	1,205	776	473	1.19	0.78	4.22	4.11	0.90	0.55	0.95	0.71
3 Odds	End	1,223	912	735	1.33	0.91	4.40	4.11	0.87	0.43	0.93	0.60
3 Evens	End-Timed	1,263	794	468	1.13	0.72	4.20	4.10	0.90	0.57	0.95	0.73
	End-Timed	1,242	965	769	1.34	0.90	4.47	4.16	0.86	0.40	0.93	0.58
	Interspersed	1,955	1,418	788	1.19	0.72	4.37	4.18	0.87	0.42	0.93	0.59
	Interspersed	1,234	1,086	824	1.43	0.96	4.53	4.27	0.83	0.39	0.91	0.56
5	End	1,208	759	482	0.92	0.54	4.74	4.77	0.82	0.23	0.90	0.37
3 Odds	End	1,180	488	312	0.57	0.33	5.04	4.86	0.84	0.66	0.91	0.79
2 Evens	End-Timed	1,302	800	516	0.90	0.52	4.75	4.83	0.83	0.24	0.90	0.38
	End-Timed	1,205	492	307	0.56	0.32	4.99	4.85	0.83	0.63	0.91	0.77
	Interspersed	1,860	1,256	809	0.97	0.56	5.04	4.93	0.81	0.40	0.90	0.57
	Interspersed	1,287	733	463	0.78	0.44	5.13	4.96	0.80	0.53	0.89	0.69
6	End	1,102	760	525	1.13	0.73	4.19	4.31	0.87	0.22	0.93	0.35
4 Odds	End-Timed	1,437	997	729	1.21	0.80	4.15	4.26	0.89	0.32	0.94	0.48
3 Evens	Interspersed	1,093	886	630	1.35	0.88	4.42	4.50	0.86	0.09	0.93	0.16

Table G-2. Split Half Data for Number Correct (NC) and Log Correct Response Time (LCT) – Students with at least 50% Correct

		Number of Cases										
		All	NC > 0		NC Means		LCT Means		Odd-Even Corr.		Reliability Estimate	
Grade	Cond.	Students	Odds	Evens	Odds	Evens	Odds	Evens	NC	LCT	NC	LCT
3	End	640	640	640	3.05	2.62	3.89	3.78	0.73	0.33	0.92	0.73
4 Odds	End-Timed	676	676	676	3.07	2.64	3.88	3.77	0.74	0.40	0.93	0.77
4 Evens	Interspersed	1,025	1,025	1,025	2.98	2.55	4.07	4.06	0.71	0.45	0.92	0.80
4	End	433	433	433	2.52	2.06	4.13	4.10	0.64	0.55	0.89	0.85
3 Odds	End	530	530	530	2.35	1.72	4.30	4.10	0.68	0.51	0.90	0.84
3 Evens	End-Timed	403	403	403	2.56	2.09	4.08	4.09	0.68	0.62	0.90	0.88
	End-Timed	519	519	519	2.34	1.67	4.38	4.14	0.71	0.47	0.91	0.81
	Interspersed	606	606	606	2.48	2.01	4.17	4.19	0.68	0.49	0.91	0.82
	Interspersed	523	523	523	2.30	1.68	4.42	4.23	0.72	0.50	0.92	0.83
5	End	309	309	309	2.15	1.53	5.02	4.98	0.40	0.46	0.77	0.81
3 Odds	End	169	169	169	2.12	1.44	4.92	4.83	0.42	0.77	0.79	0.94
2 Evens	End-Timed	314	314	314	2.17	1.50	5.05	4.98	0.45	0.47	0.81	0.81
	End-Timed	163	163	163	2.13	1.47	4.89	4.80	0.41	0.71	0.78	0.92
	Interspersed	467	467	467	2.18	1.51	5.07	4.94	0.46	0.54	0.81	0.85
	Interspersed	250	250	250	2.09	1.41	5.01	4.95	0.38	0.52	0.76	0.84
6	End	225	225	225	2.60	2.23	4.42	4.36	0.61	0.31	0.88	0.71
4 Odds	End-Timed	310	310	310	2.77	2.34	4.35	4.29	0.69	0.46	0.91	0.81
3 Evens	Interspersed	262	262	262	2.67	2.26	4.66	4.58	0.56	0.12	0.86	0.52

Appendix H. Original vs. Alternative Scoring Rules

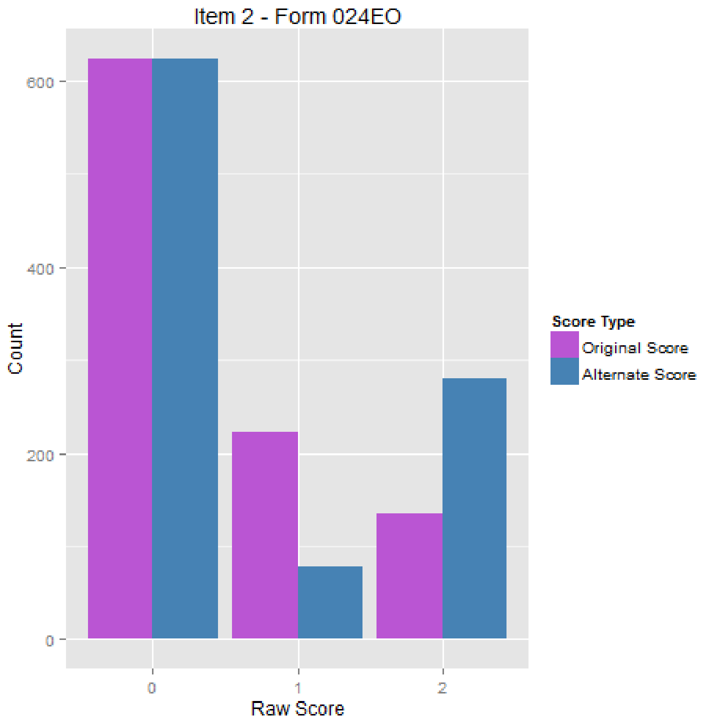


Figure H-1. The number of individuals who receive no credit, partial credit, and full credit for Item 2 - Form 024EO using the original and alternate scoring rules.

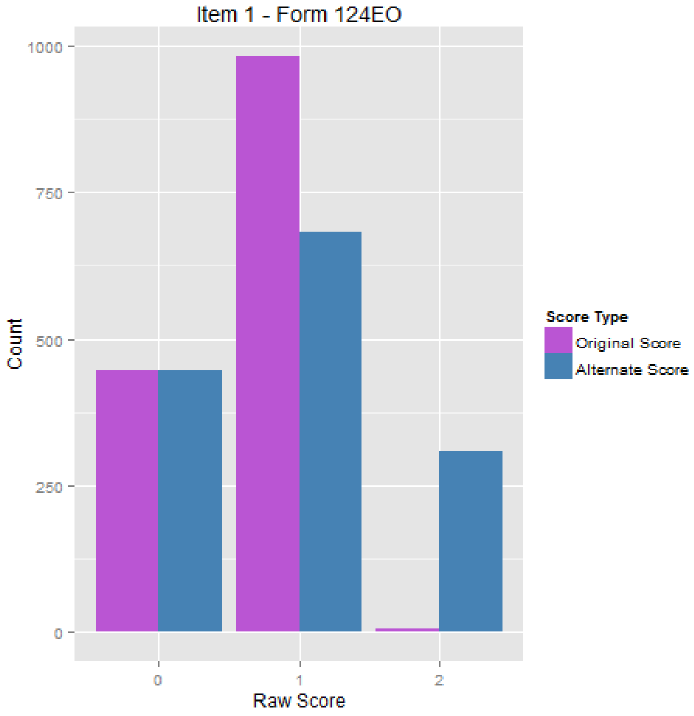


Figure H-2. The number of individuals who receive no credit, partial credit, and full credit for Item 1 – Form 124EO using the original and alternate scoring rules.

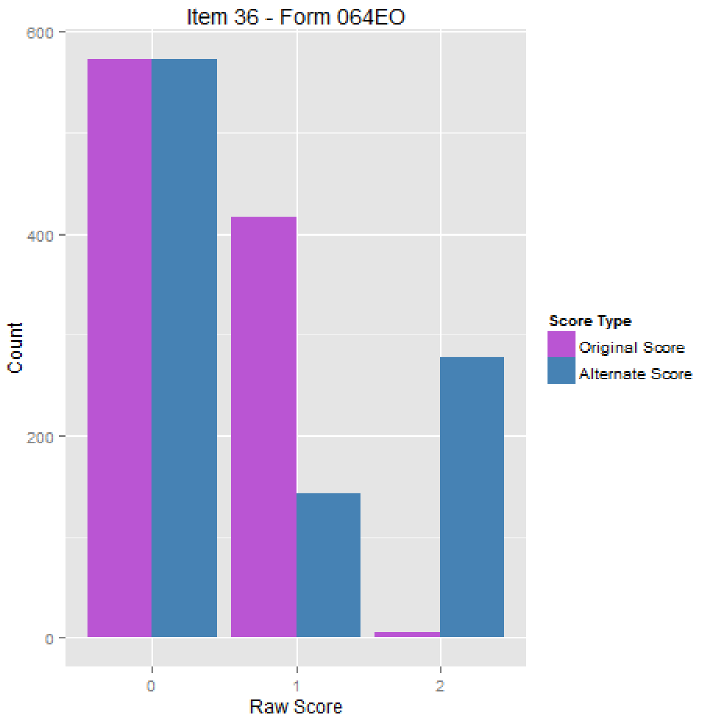


Figure H-3. The number of individuals who receive no credit, partial credit, and full credit for Item 36 – Form 064EO using the original and alternate scoring rules.

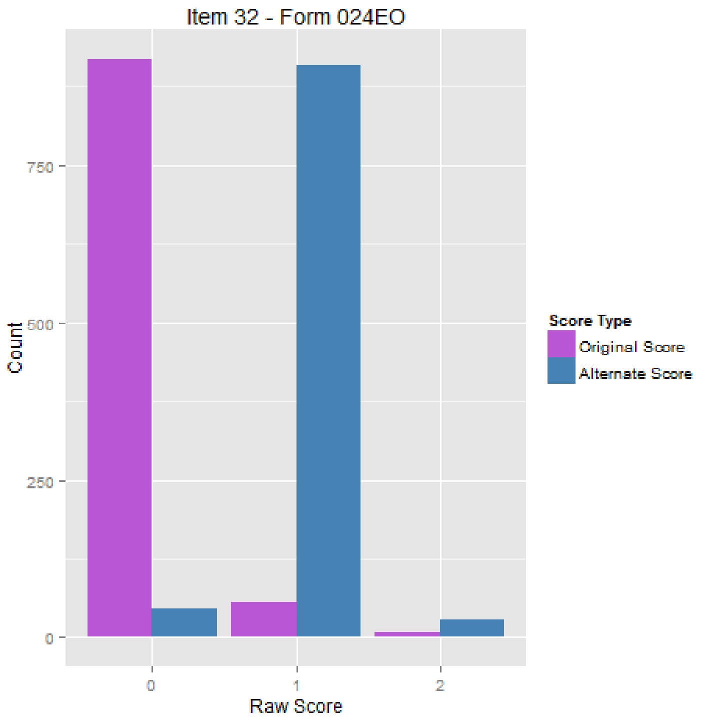


Figure H-4. The number of individuals who receive no credit, partial credit, and full credit for Item 32 – Form 024EO using the original and alternate scoring rules.

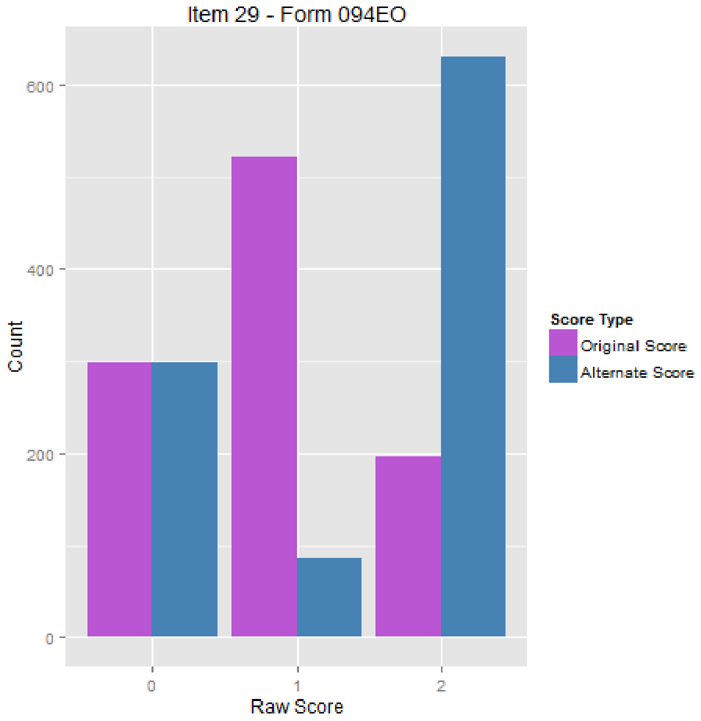


Figure H-5. The number of individuals who receive no credit, partial credit, and full credit for Item 29 – Form 094EO using the original and alternate scoring rules.

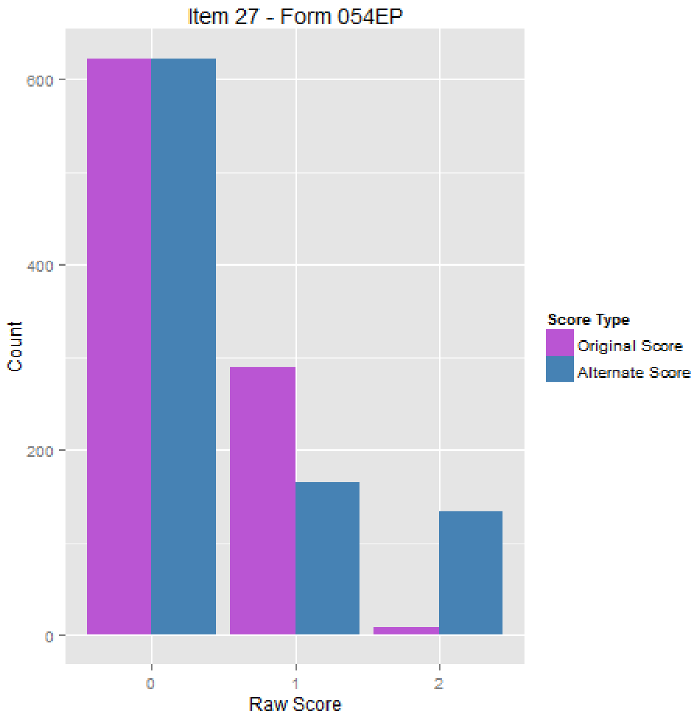


Figure H-6. The number of individuals who receive no credit, partial credit, and full credit for Item 27 – Form 054EP using the original and alternate scoring rules.

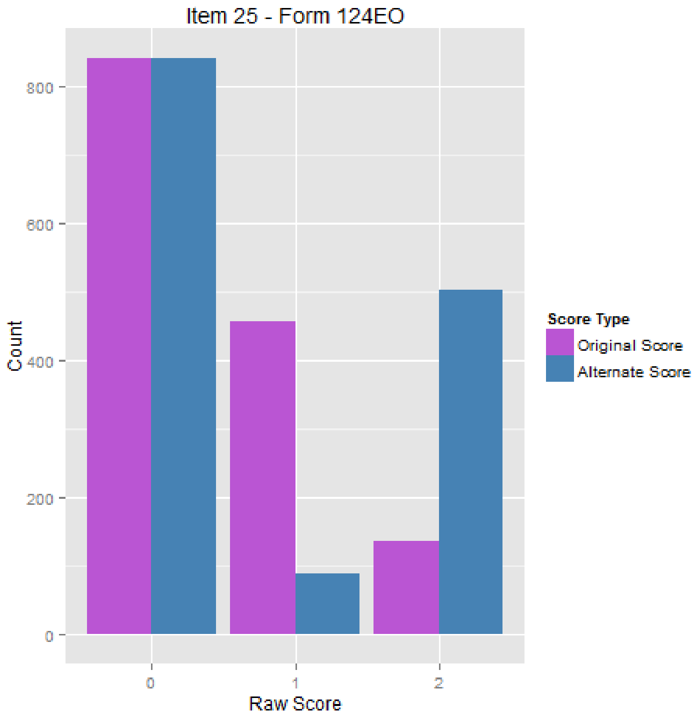


Figure H-7. The number of individuals who receive no credit, partial credit, and full credit for Item 25 – Form 124EO using the original and alternate scoring rules.

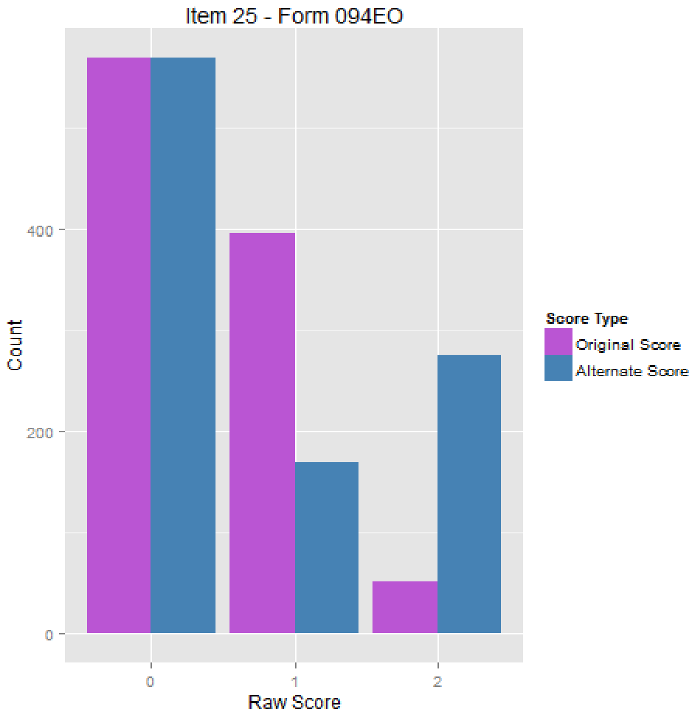


Figure H-8. The number of individuals who receive no credit, partial credit, and full credit for Item 25 – Form 094EO using the original and alternate scoring rules.

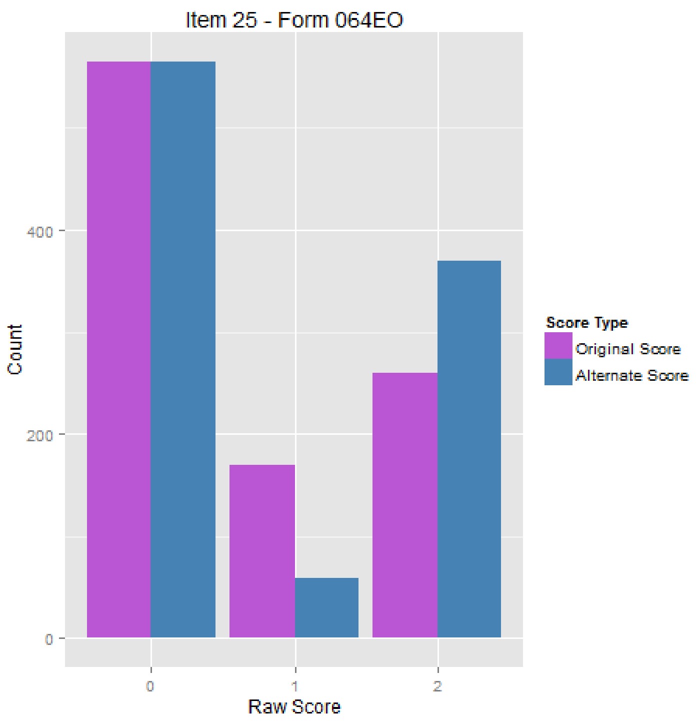


Figure H-9. The number of individuals who receive no credit, partial credit, and full credit for Item 25 – Form 064EO using the original and alternate scoring rules.

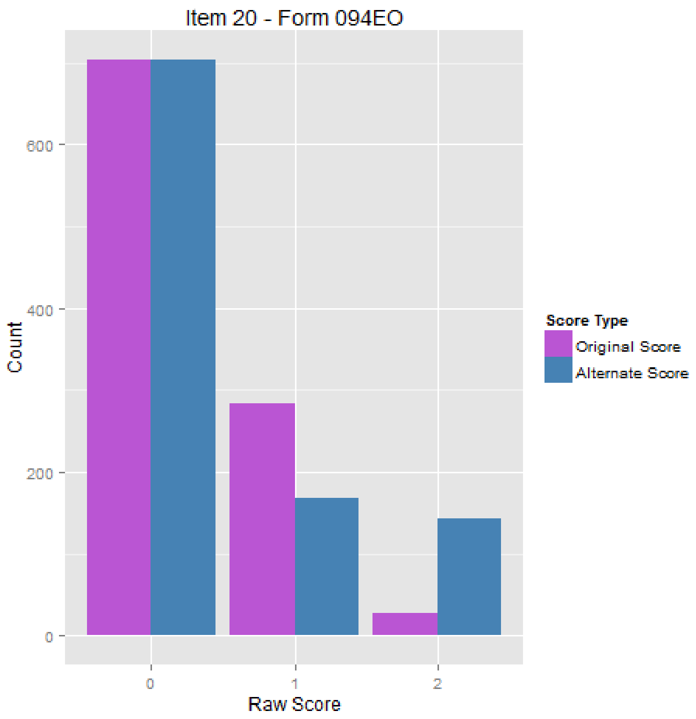


Figure H-10. The number of individuals who receive no credit, partial credit, and full credit for Item 20 – Form 094EO using the original and alternate scoring rules.

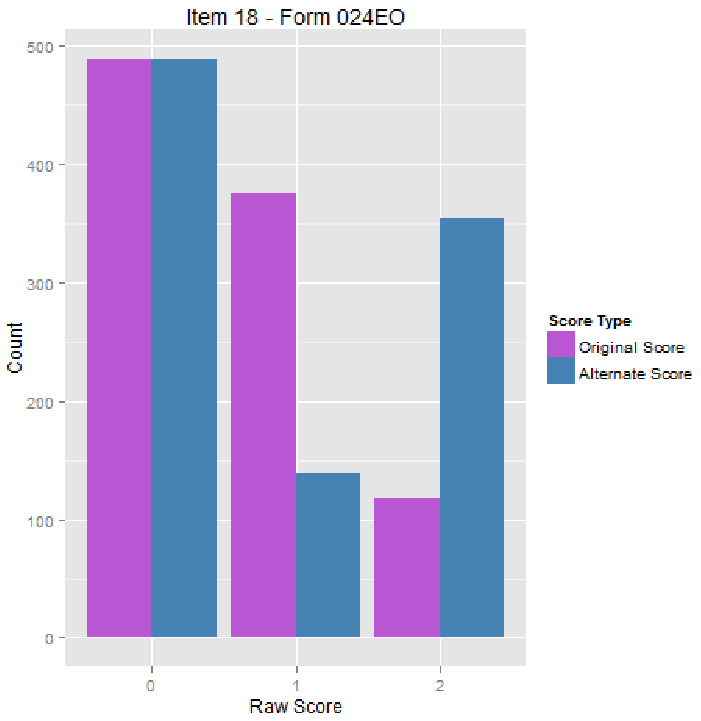


Figure H-11. The number of individuals who receive no credit, partial credit, and full credit for Item 18 – Form 024EO using the original and alternate scoring rules.

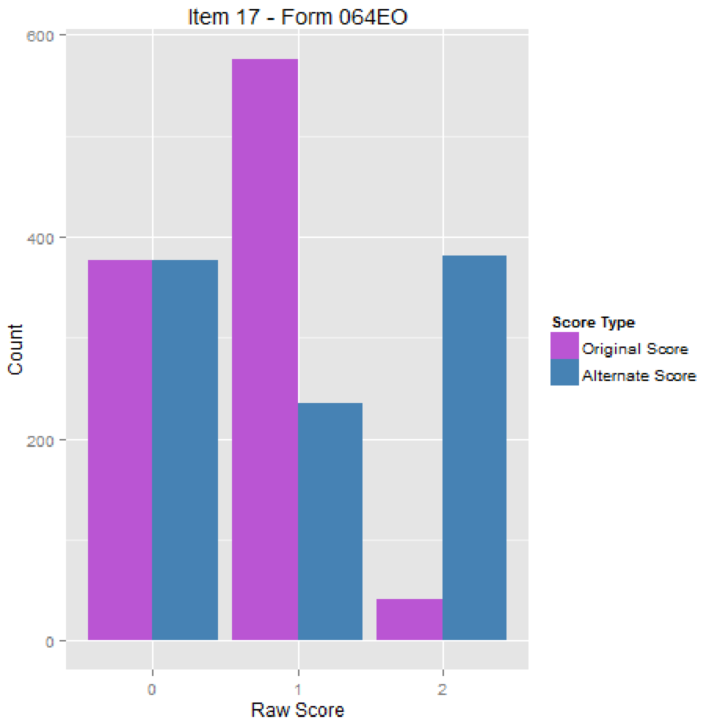


Figure H-12. The number of individuals who receive no credit, partial credit, and full credit for Item 17 – Form 064EO using the original and alternate scoring rules.

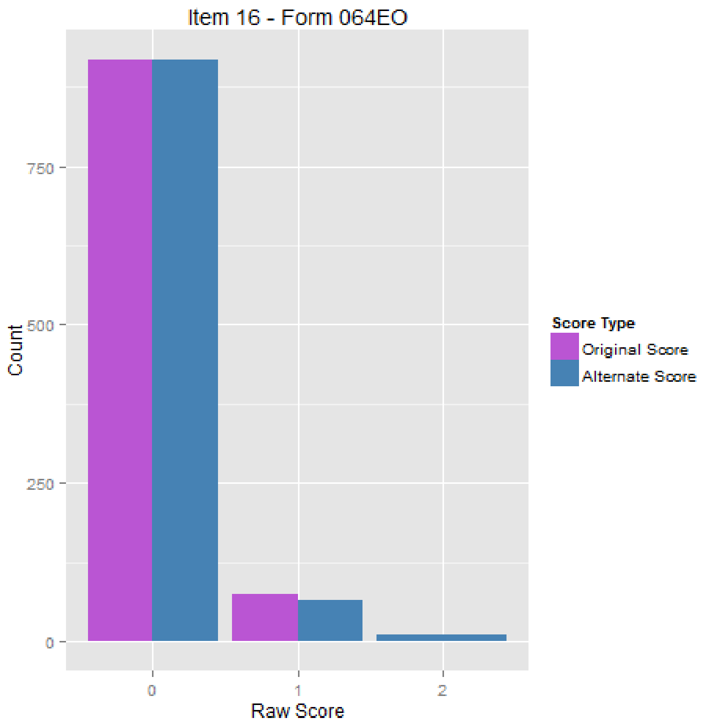


Figure H-13. The number of individuals who receive no credit, partial credit, and full credit for Item16 – Form 064EO using the original and alternate scoring rules.

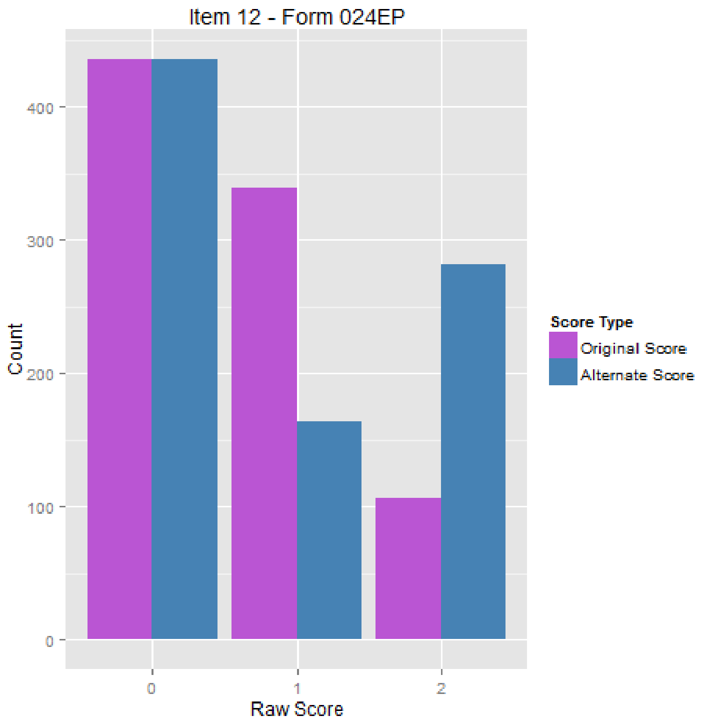


Figure H-14. The number of individuals who receive no credit, partial credit, and full credit for Item 12 – Form 024EP using the original and alternate scoring rules.

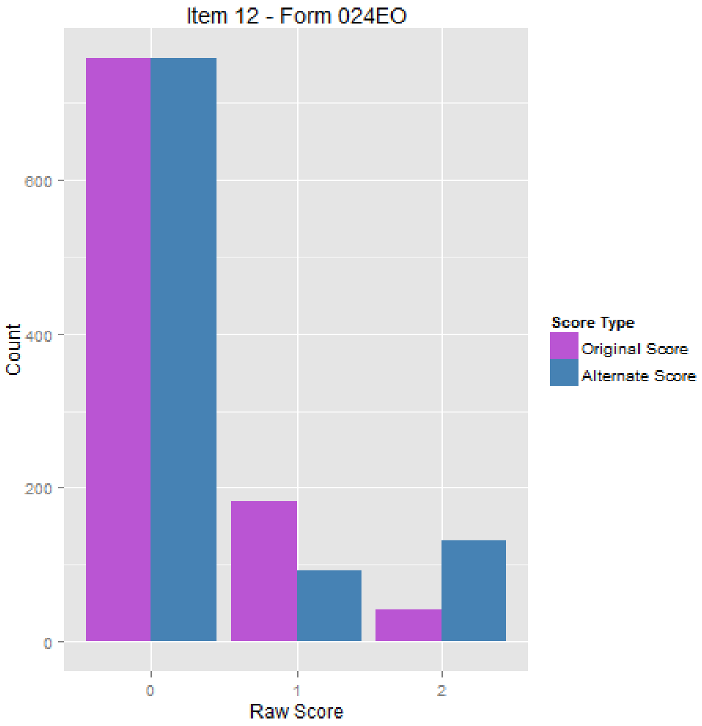


Figure H-15. The number of individuals who receive no credit, partial credit, and full credit for Item 12 – Form 024EO using the original and alternate scoring rules.

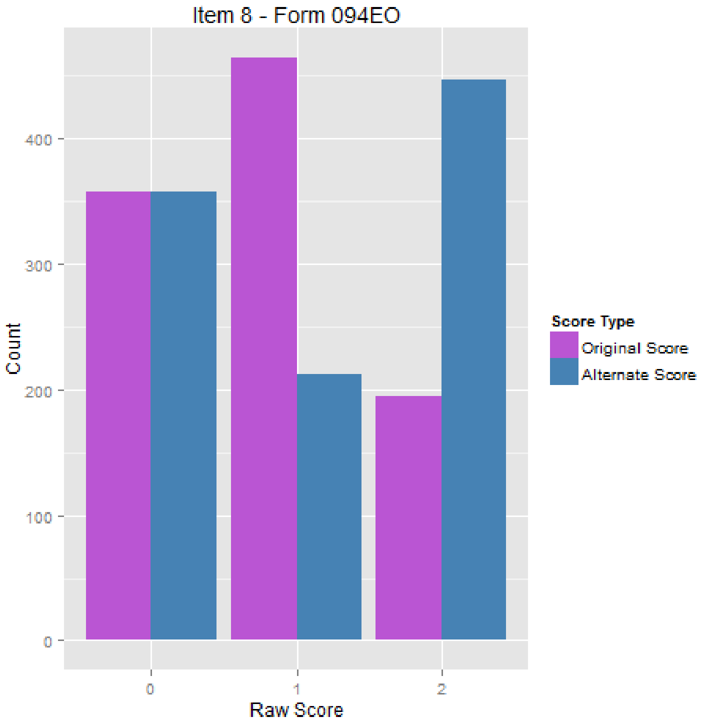


Figure H-16. The number of individuals who receive no credit, partial credit, and full credit for Item 8 – Form 094EO using the original and alternate scoring rules.

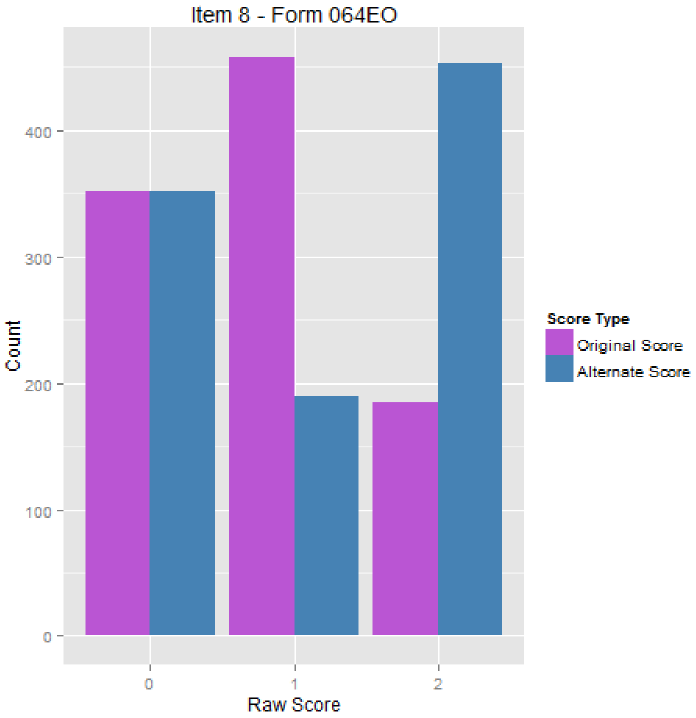


Figure H-17. The number of individuals who receive no credit, partial credit, and full credit for Item 8 – Form 064EO using the original and alternate scoring rules.

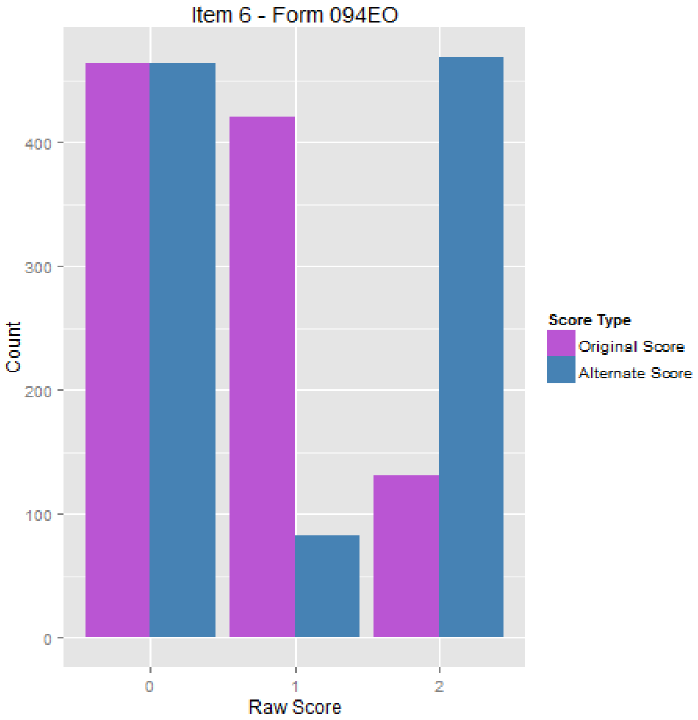


Figure H-18. The number of individuals who receive no credit, partial credit, and full credit for Item 6 – Form 094EO using the original and alternate scoring rules.

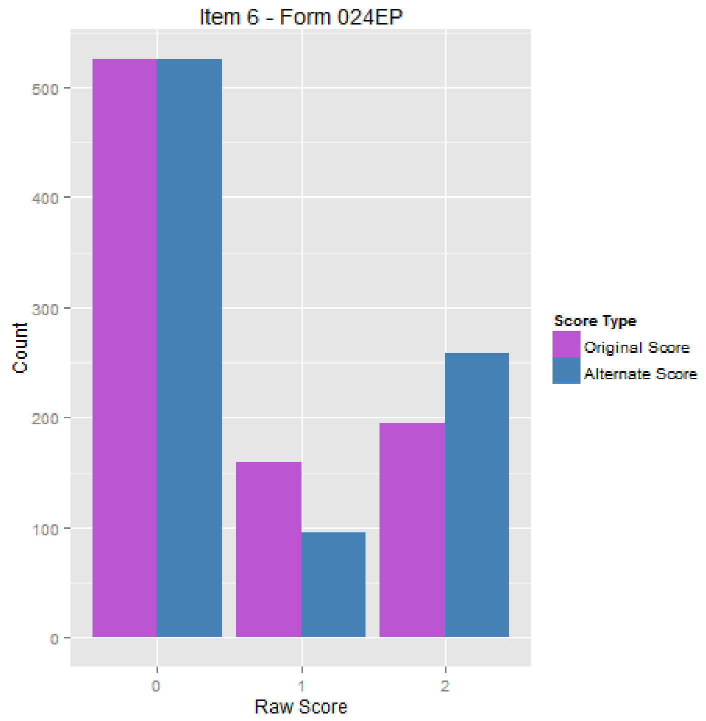


Figure H-19. The number of individuals who receive no credit, partial credit, and full credit for Item 6 – Form 024EP using the original and alternate scoring rules.

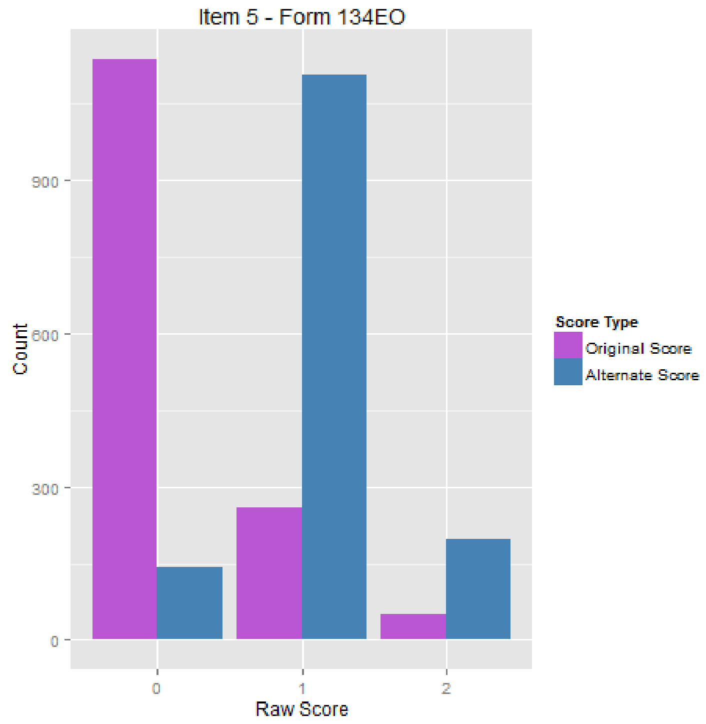


Figure H-20. The number of individuals who receive no credit, partial credit, and full credit for Item 5 – Form 134EO using the original and alternate scoring rules.

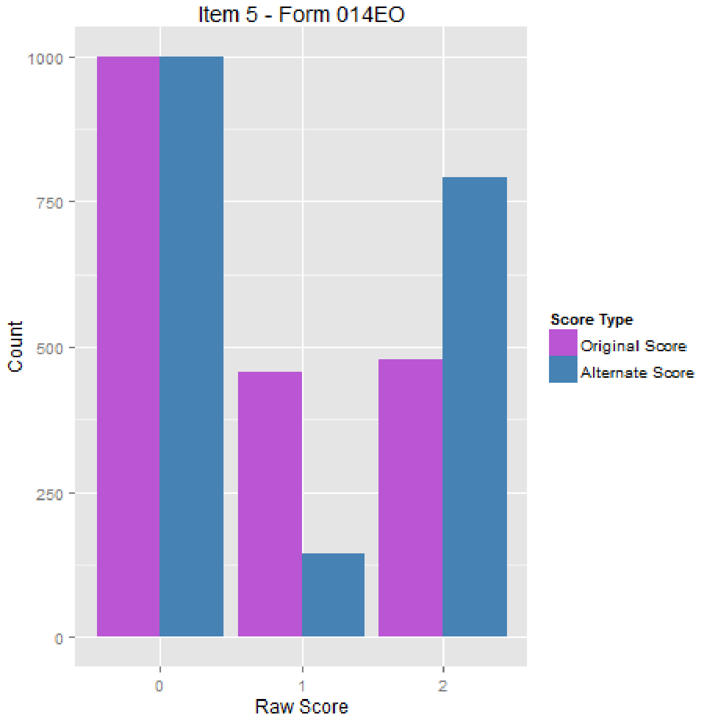


Figure H-21. The number of individuals who receive no credit, partial credit, and full credit for Item 5 – Form 014EO using the original and alternate scoring rules.

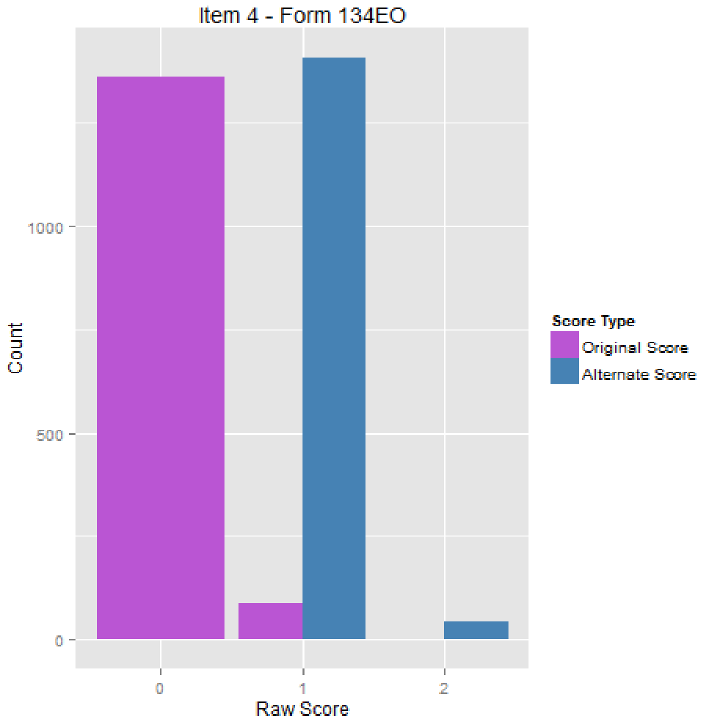


Figure H-22. The number of individuals who receive no credit, partial credit, and full credit for Item 4 – Form 134EO using the original and alternate scoring rules.

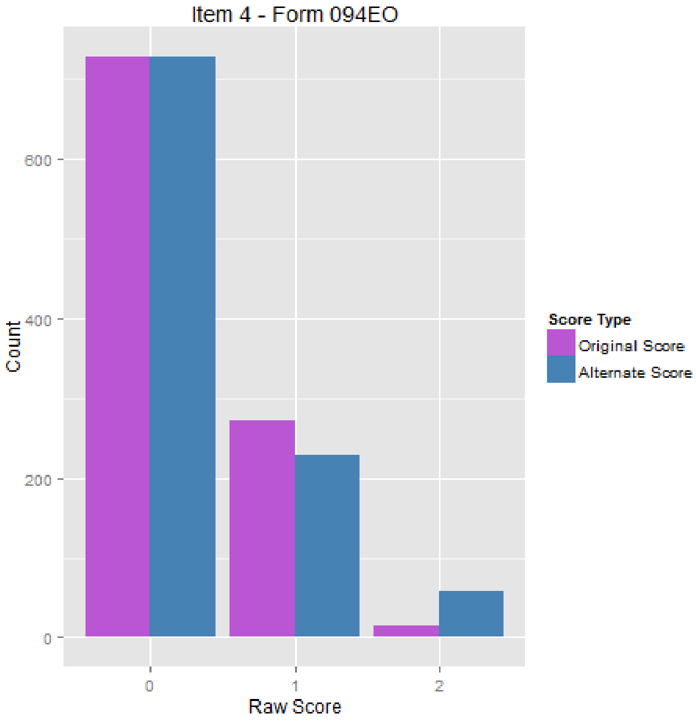


Figure H-23. The number of individuals who receive no credit, partial credit, and full credit for Item 4 – Form 094EO using the original and alternate scoring rules.

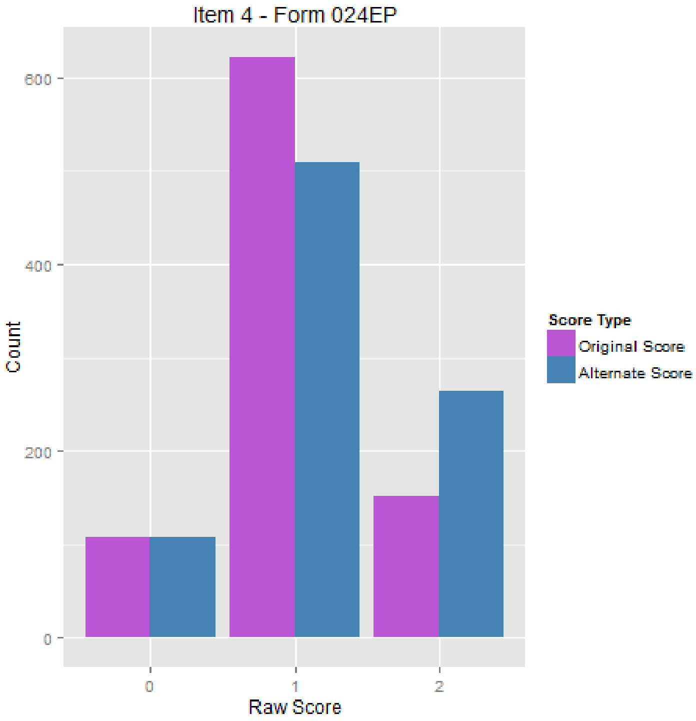


Figure H-24. The number of individuals who receive no credit, partial credit, and full credit for Item 4 – Form 024EP using the original and alternate scoring rules.

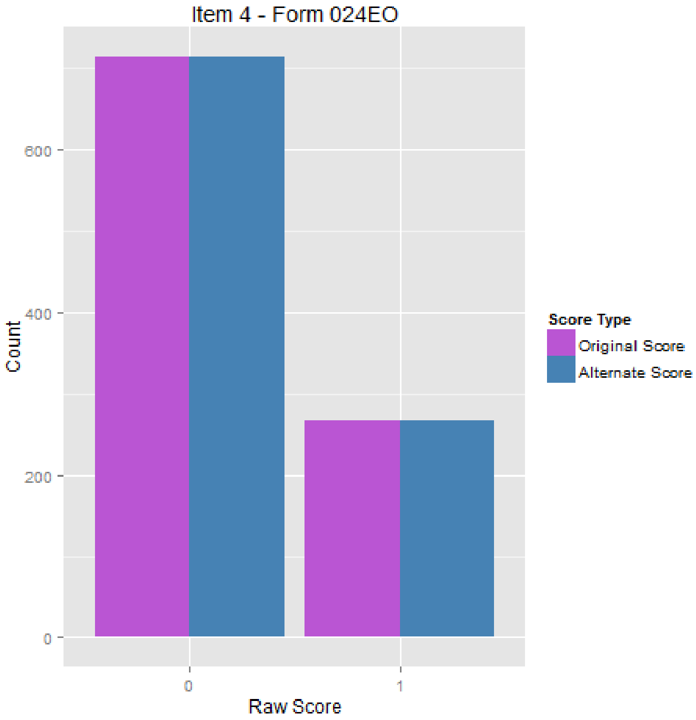
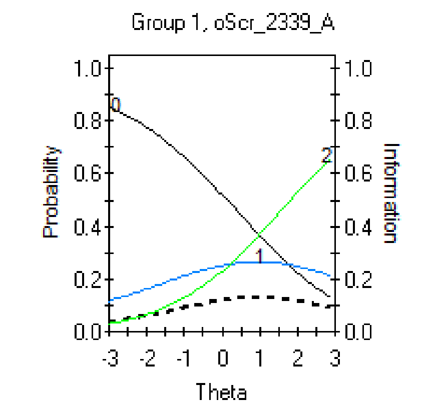
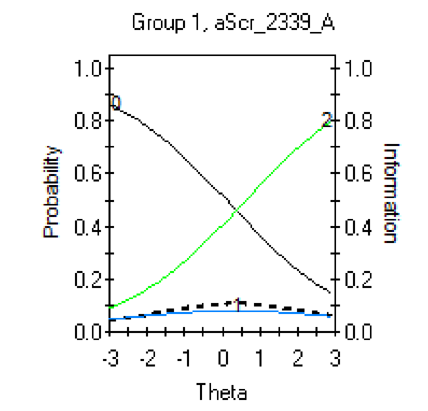
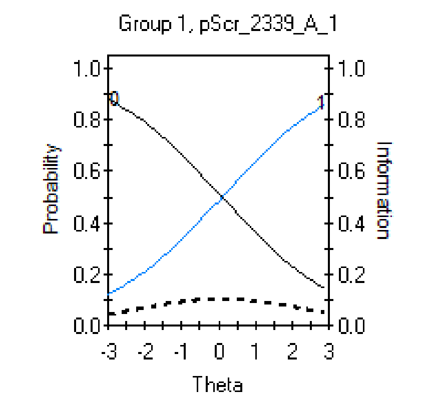
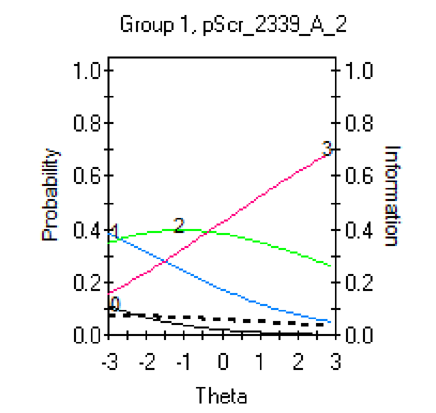
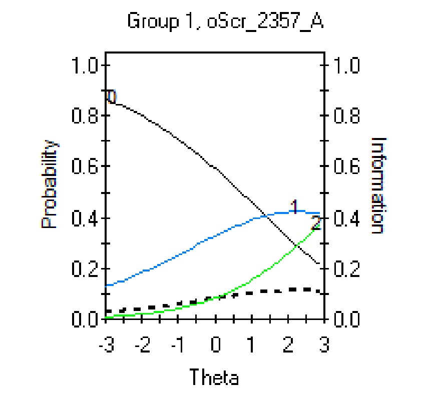
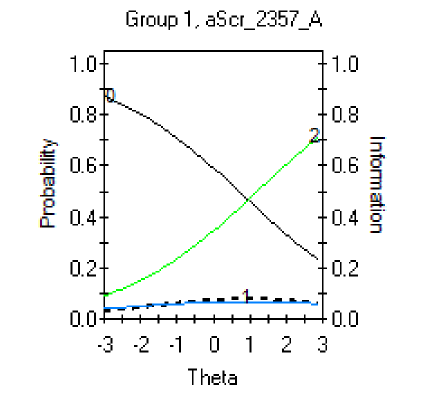
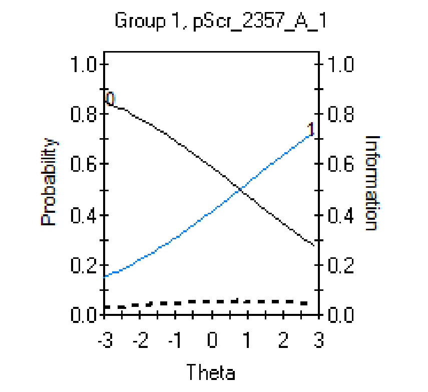
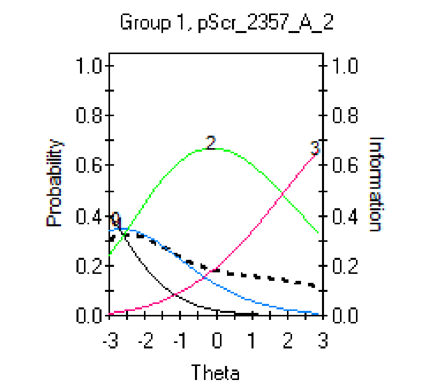
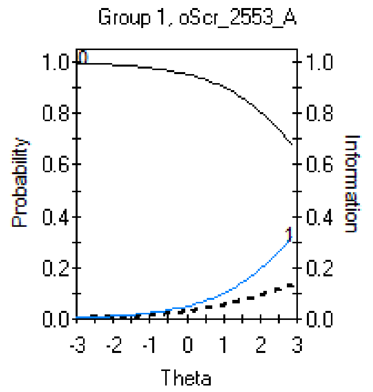
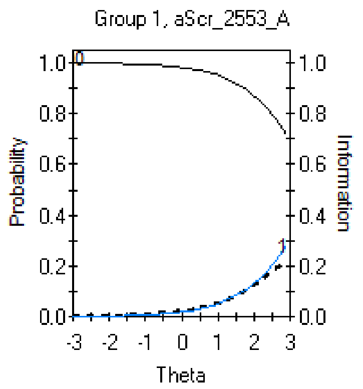
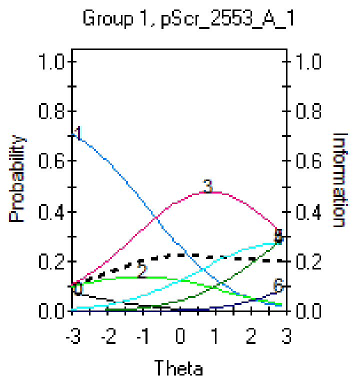
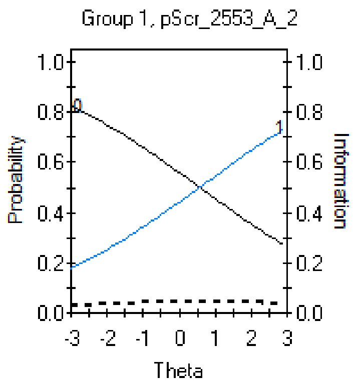
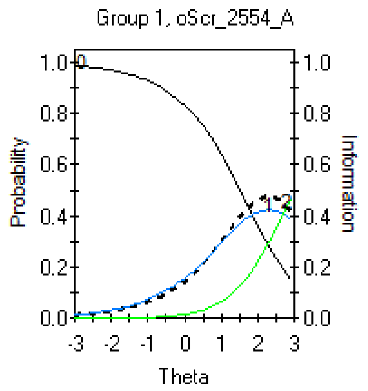
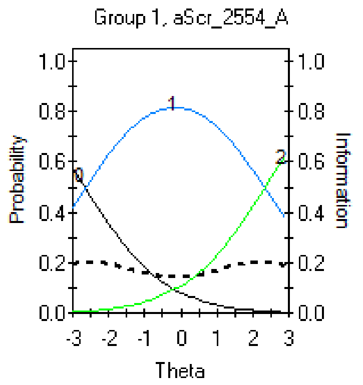
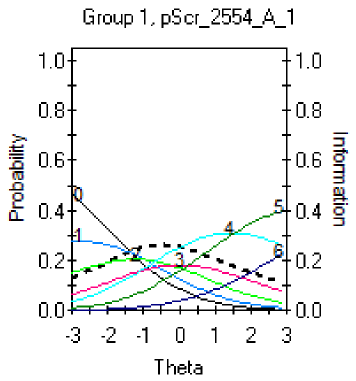
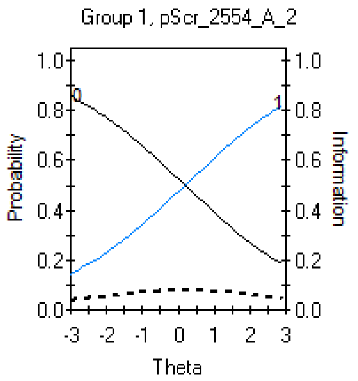


Figure H-25. The number of individuals who receive no credit, partial credit, and full credit for Item 4 – Form 024EO using the original and alternate scoring rules.

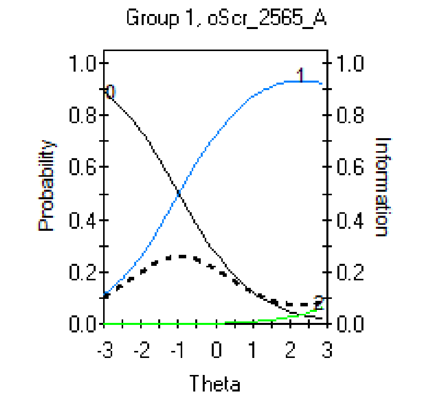
Appendix I. Item Function of Original, Alternate, and Part Scores

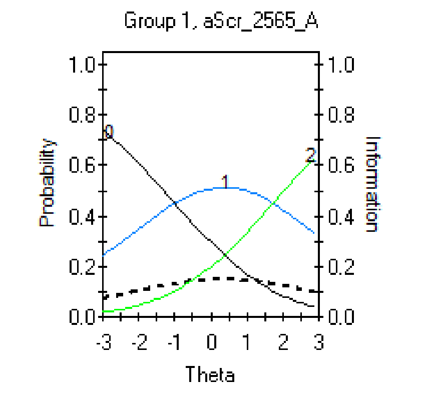
Grade 5 Items

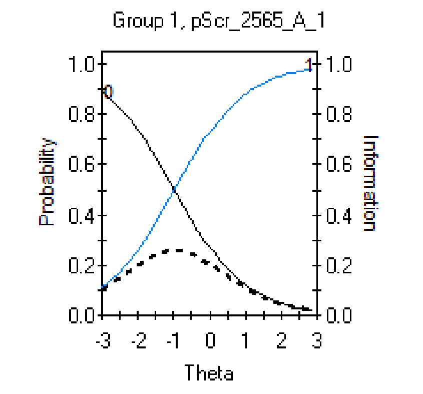
	Original Score	Alternate Score	Part 1	Part 2
014EO Item 5				
124EO Item 25				

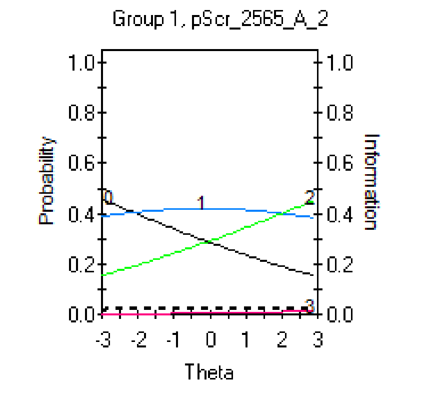
134EO Item 4	 Group 1, oScr_2553_A	 Group 1, aScr_2553_A	 Group 1, pScr_2553_A_1	 Group 1, pScr_2553_A_2
134EO Item 5	 Group 1, oScr_2554_A	 Group 1, aScr_2554_A	 Group 1, pScr_2554_A_1	 Group 1, pScr_2554_A_2

124EO Item 1

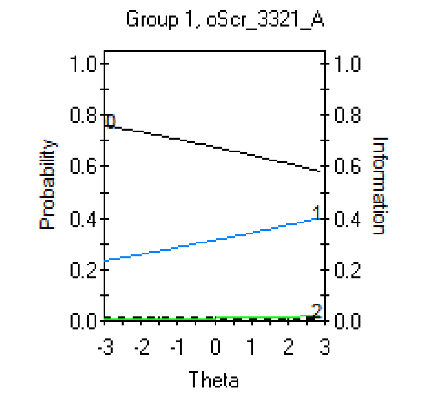


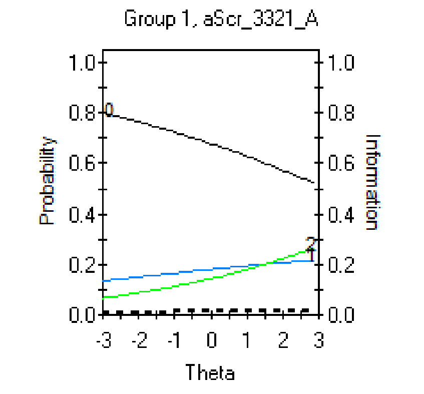


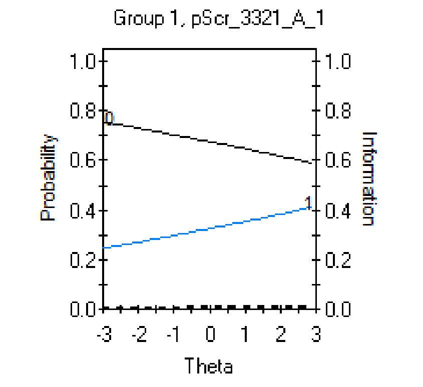


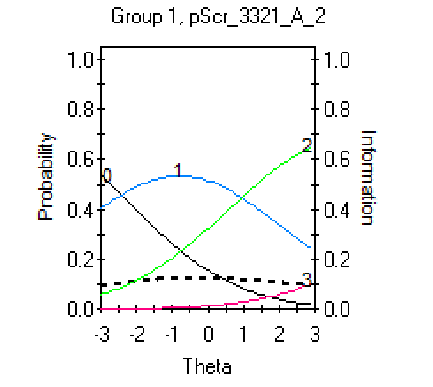


054EO Item 27

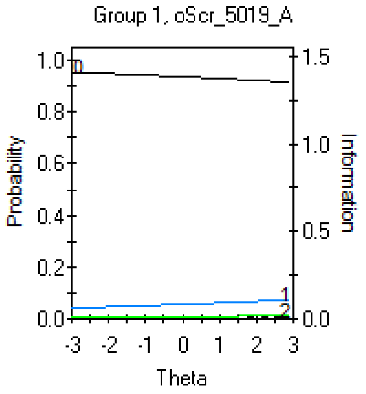
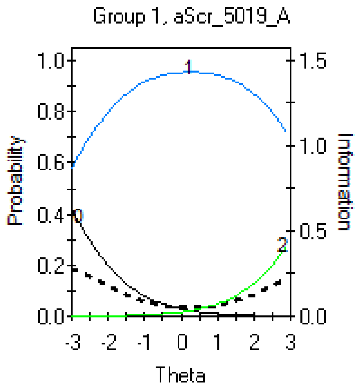
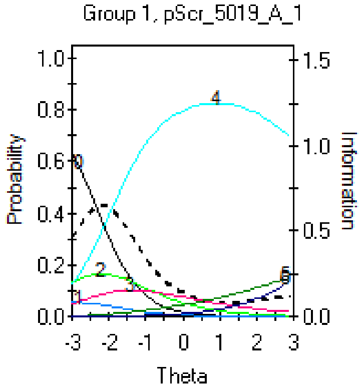
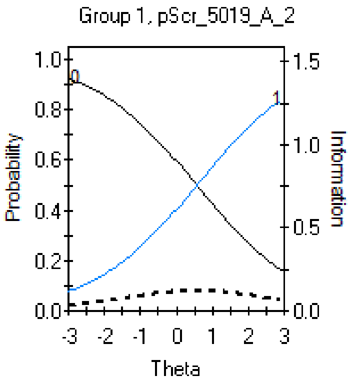
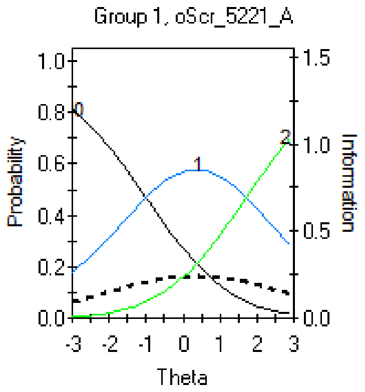
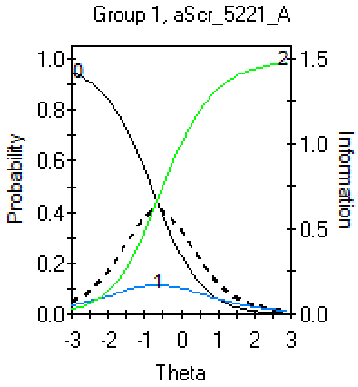
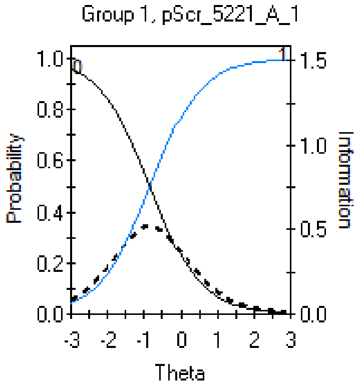
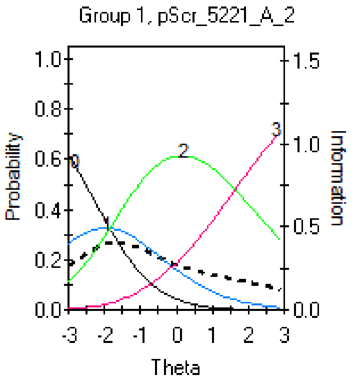


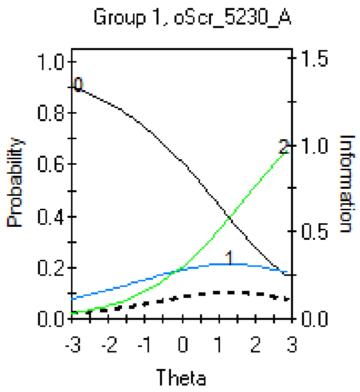
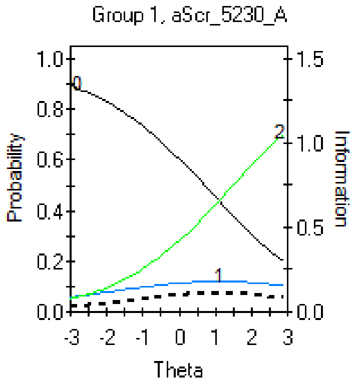
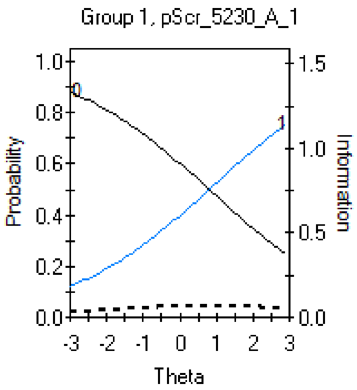
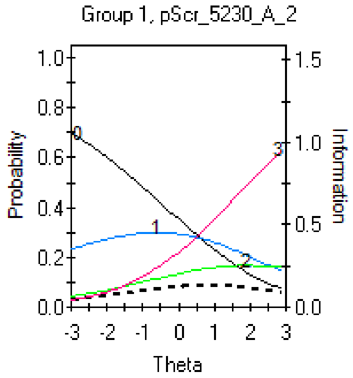
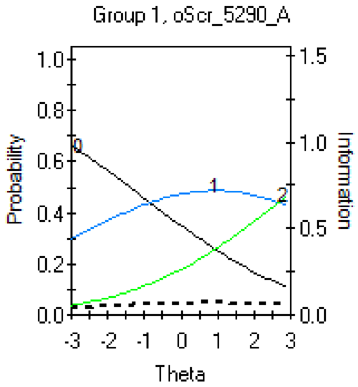
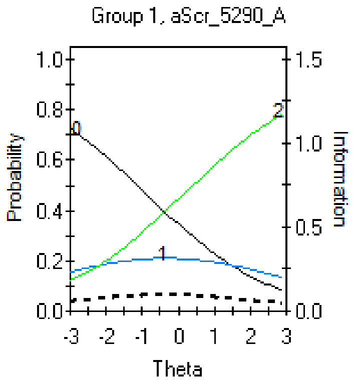
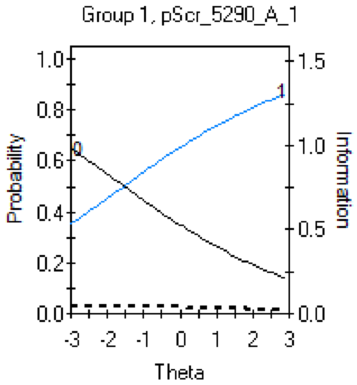
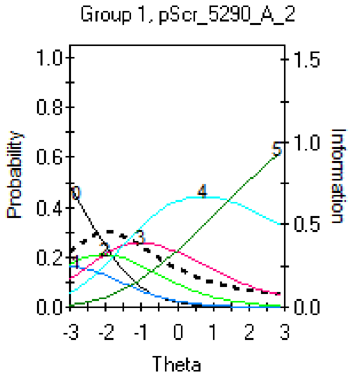


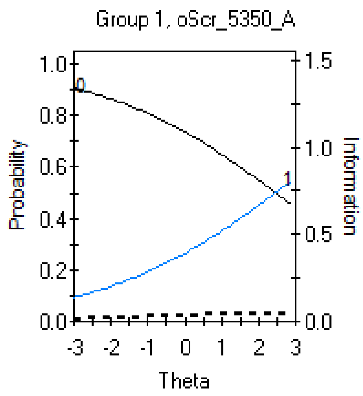
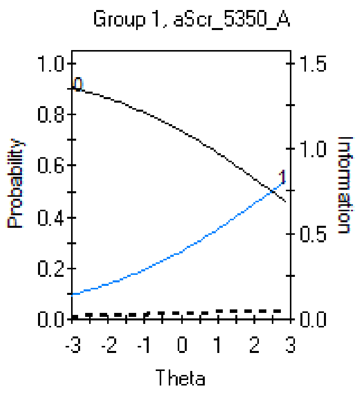
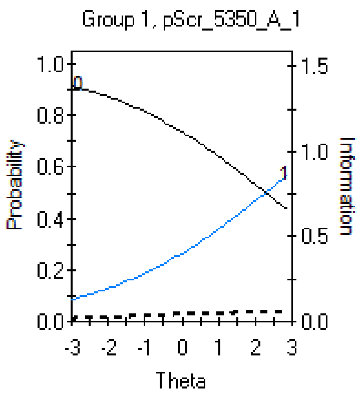
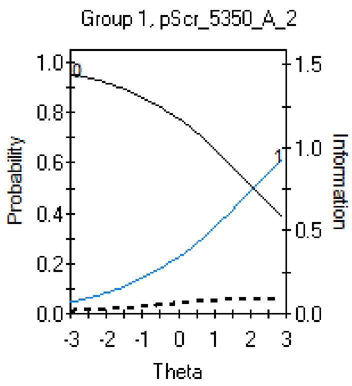
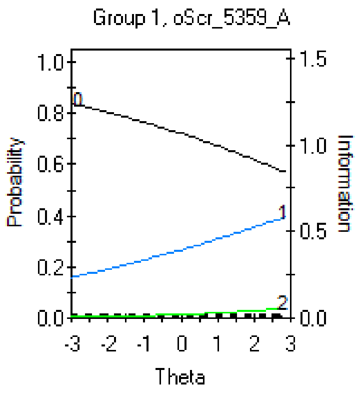
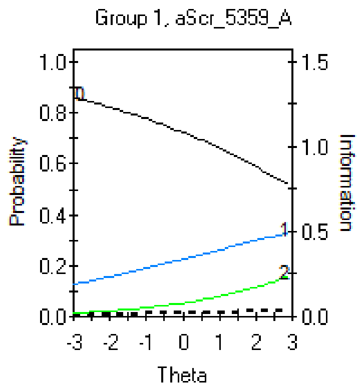
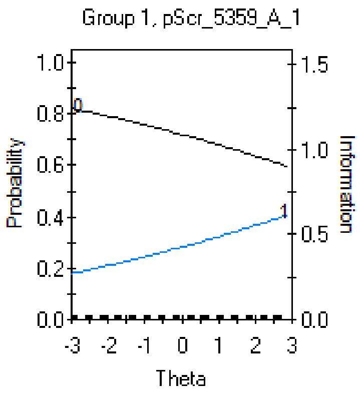
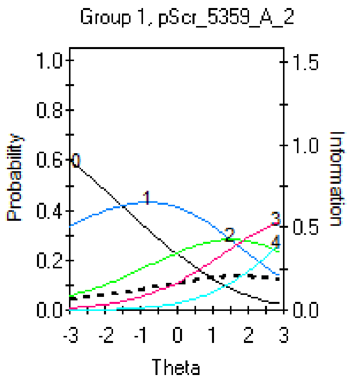


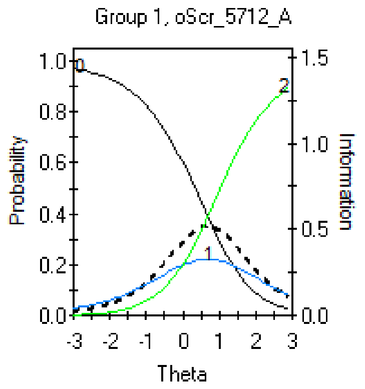
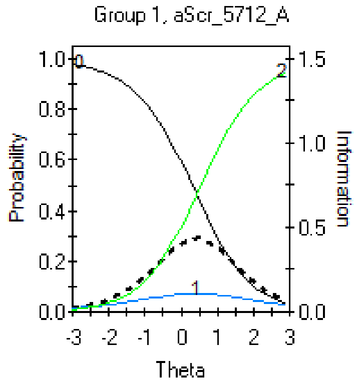
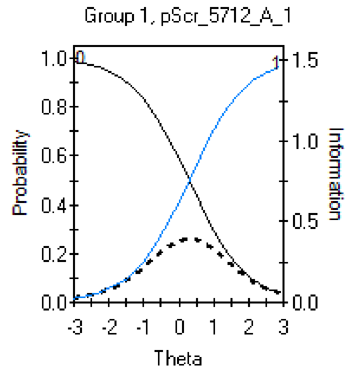
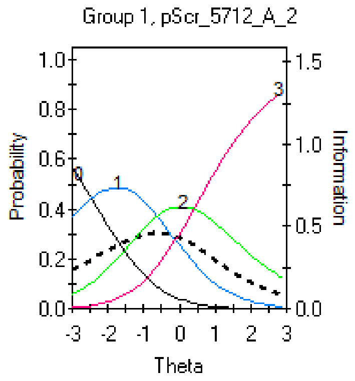
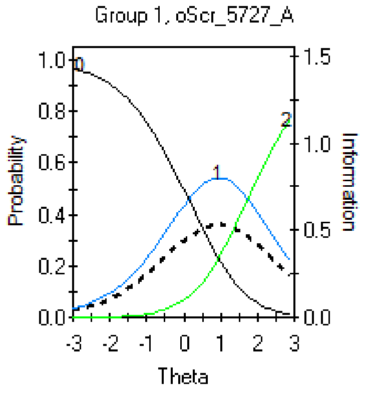
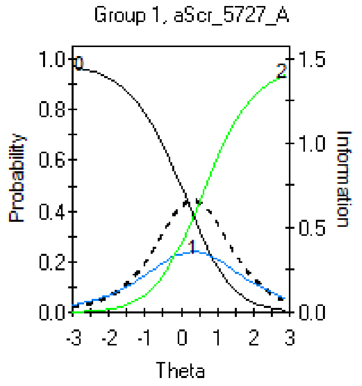
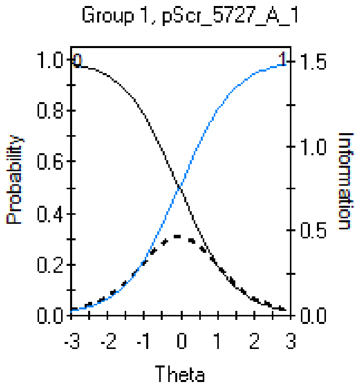
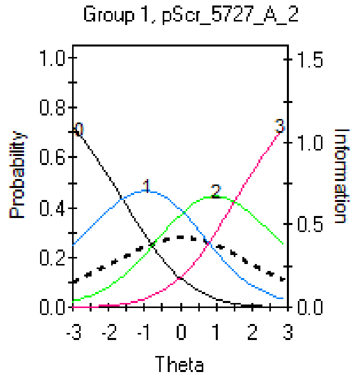


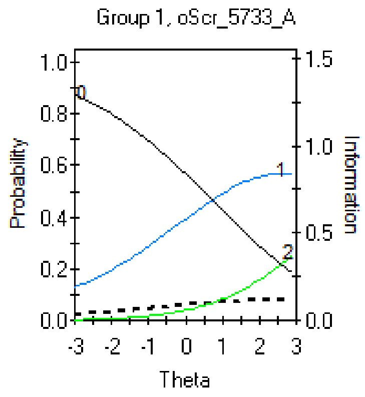
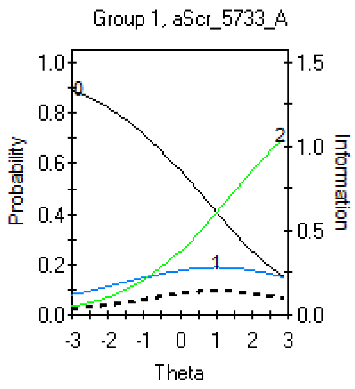
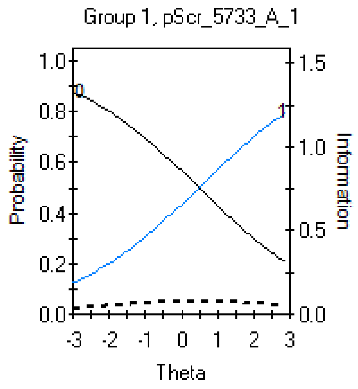
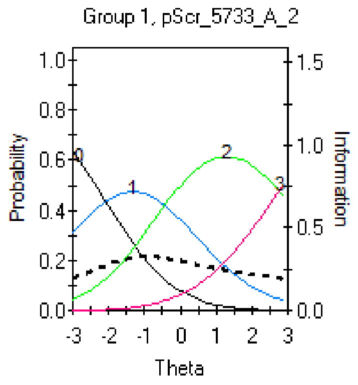
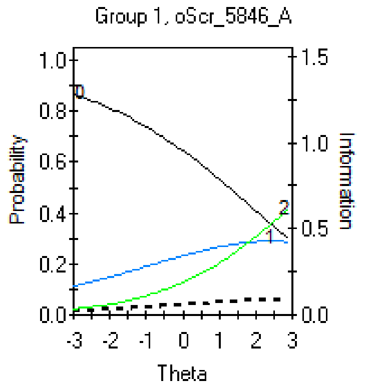
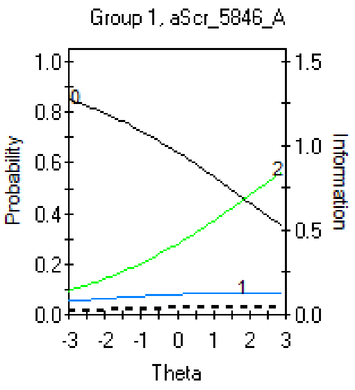
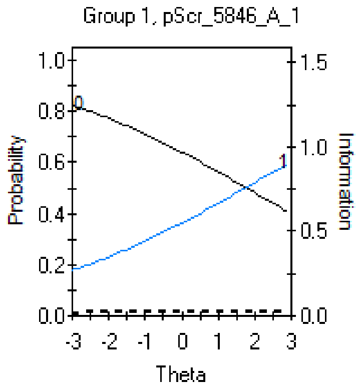
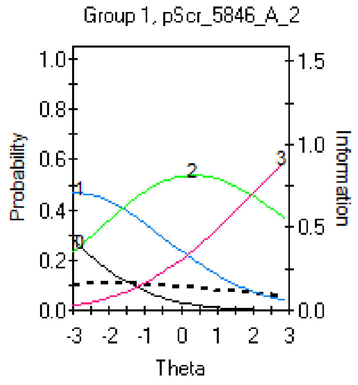
Grade 8 Items

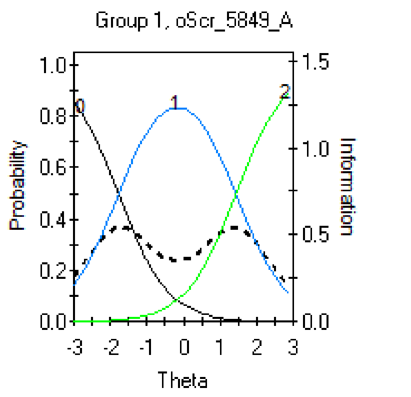
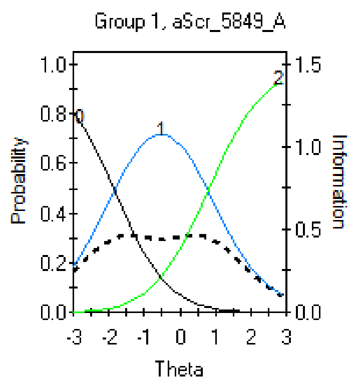
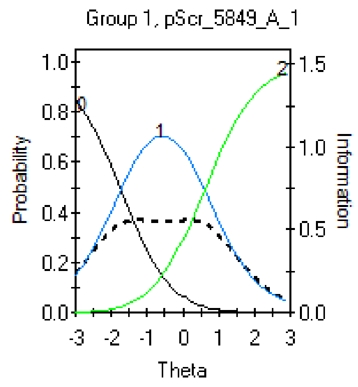
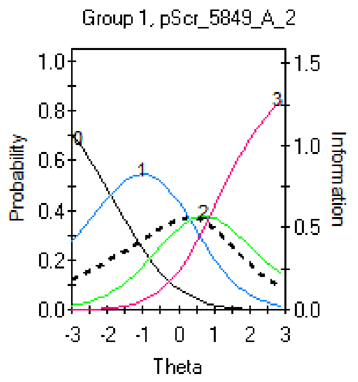
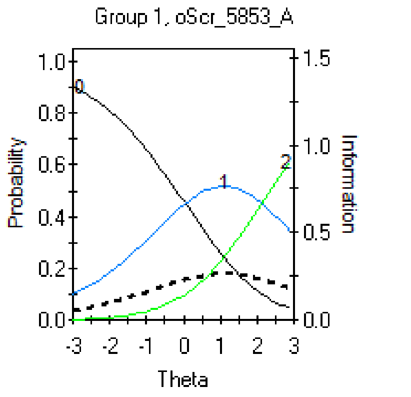
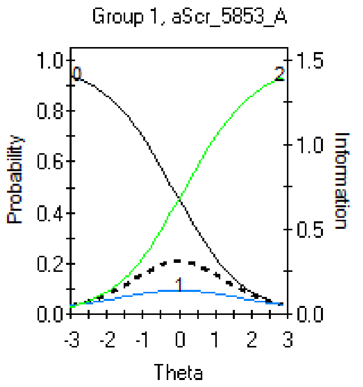
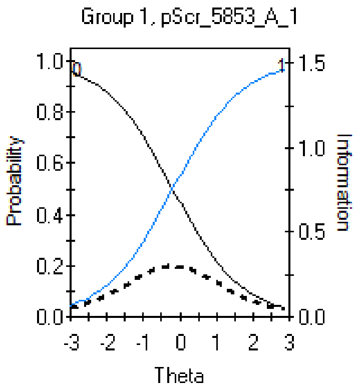
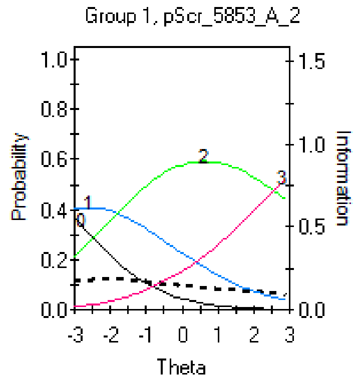
	Original Score	Alternate Score	Part 1	Part 2
024EO Item 32				
09EO Item 29				

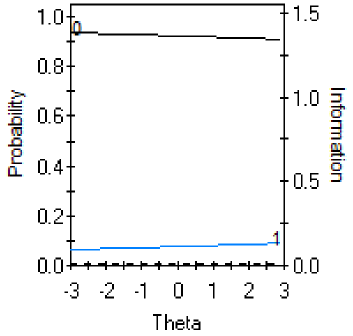
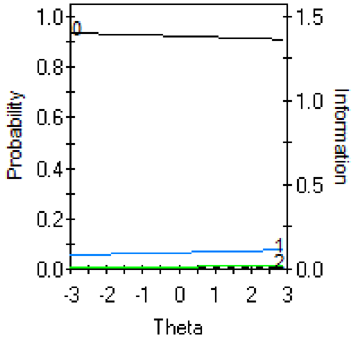
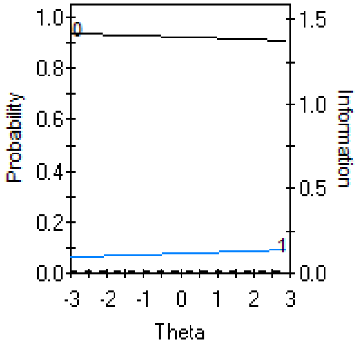
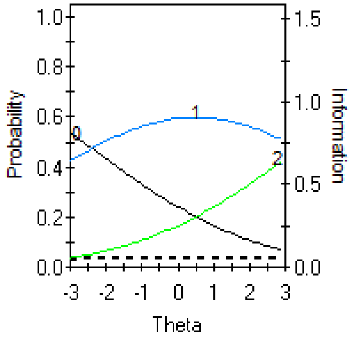
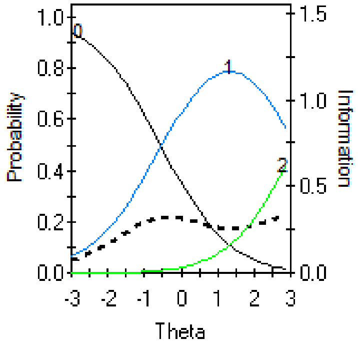
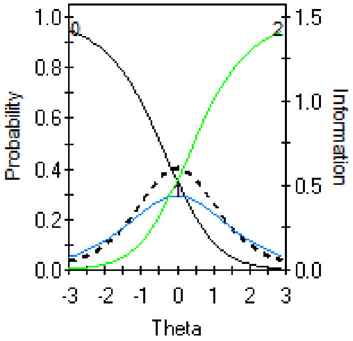
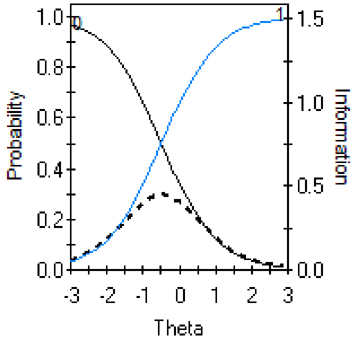
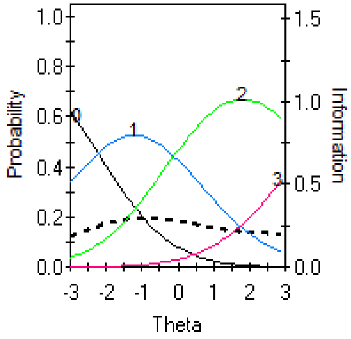
024EP Item 6				
064EO Item 8; 094EO Item 8				

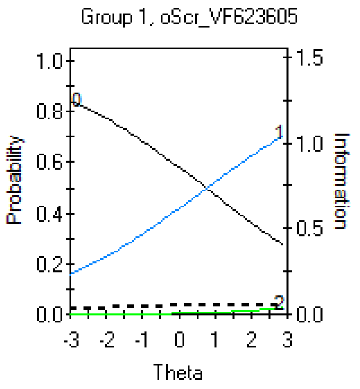
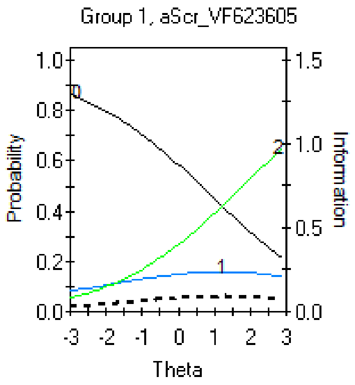
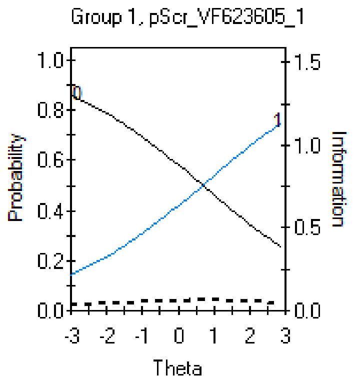
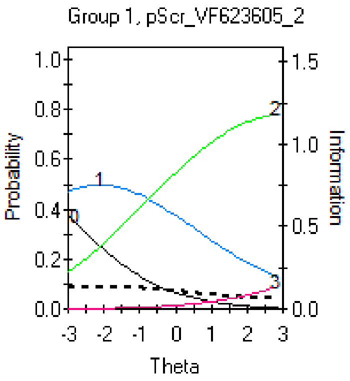
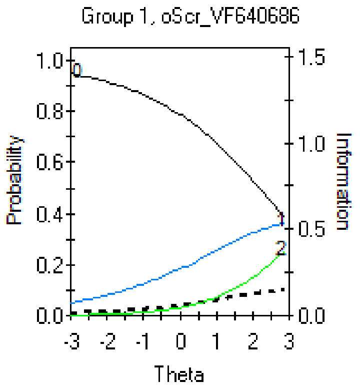
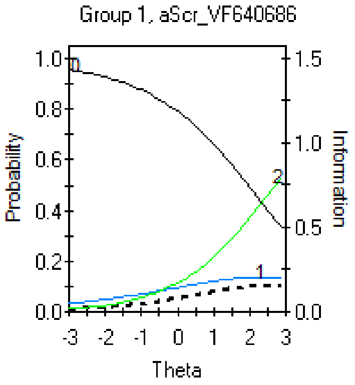
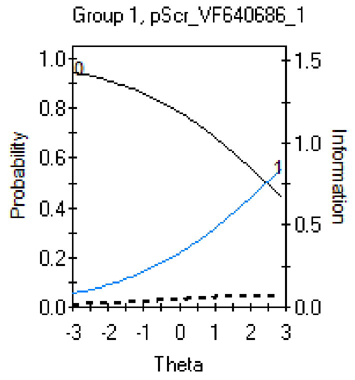
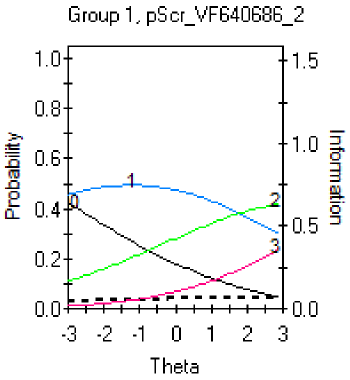
024EO Item 4	 	 	 	
094EO Item 4	 	 	 	

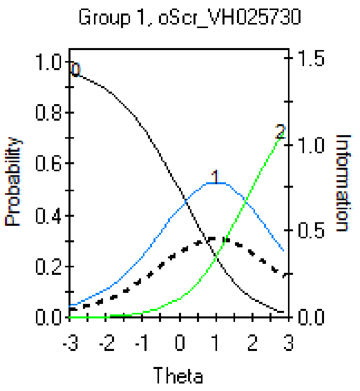
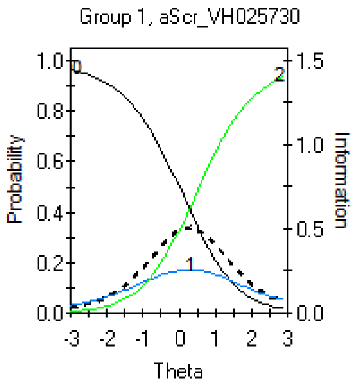
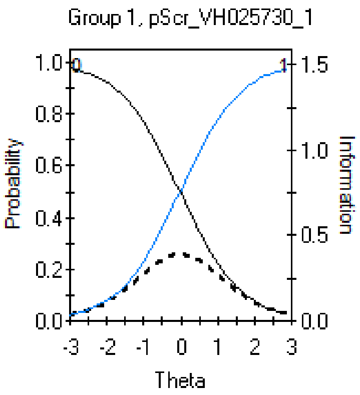
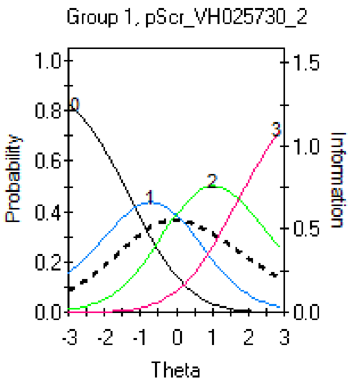
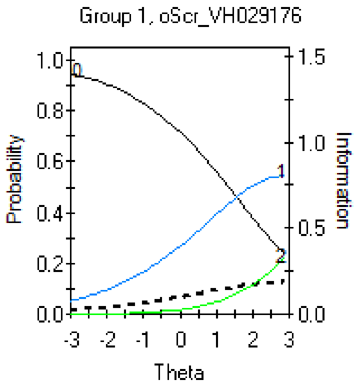
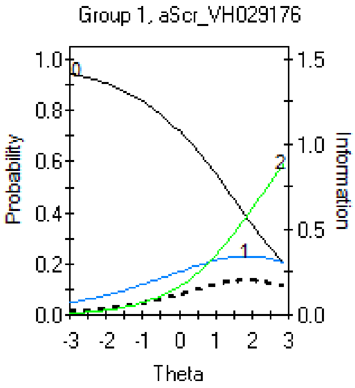
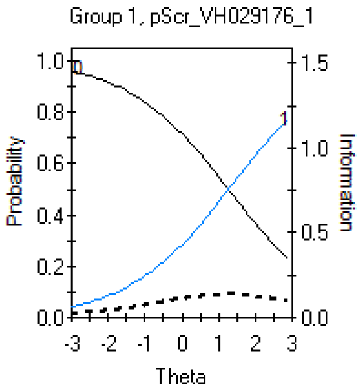
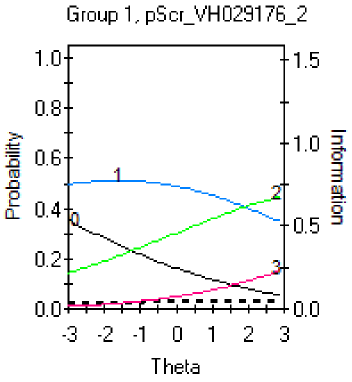
064EO Item 25	 Group 1, oScr_5712_A	 Group 1, aScr_5712_A	 Group 1, pScr_5712_A_1	 Group 1, pScr_5712_A_2
024EP Item 12	 Group 1, oScr_5727_A	 Group 1, aScr_5727_A	 Group 1, pScr_5727_A_1	 Group 1, pScr_5727_A_2

094EO Item 25				
024EO Item 2				

024EP Item 4				
094EO Item 6				

064EO Item 16	Group 1, oScr_VF560140  	Group 1, aScr_VF560140  	Group 1, pScr_VF560140_1  	Group 1, pScr_VF560140_2 
064EO Item 17	Group 1, oScr_VF560155  	Group 1, aScr_VF560155  	Group 1, pScr_VF560155_1  	Group 1, pScr_VF560155_2 

064EO Item 36				
024EO Item 12				

024EO Item 18				
09EO Item 20	 Group 1, oScr_VH029176 Probability Information Theta 0 1 2	 Group 1, aScr_VH029176 Probability Information Theta 0 1 2	 Group 1, pScr_VH029176_1 Probability Information Theta 0 1	 Group 1, pScr_VH029176_2 Probability Information Theta 0 1 2 3