KAP Technical Manual: 2015

University of Kansas Achievement & Assessment Institute

April 15, 2016

Contents

Foreword

The University of Kansas Achievement and Assessment Institute (KU AAI) is committed to following the *Standards for Educational and Psychological Testing*[1] in its work with the Kansas Assessment Program (KAP).

[1] AERA, APA, & NCME (2014)

In spring of 2015, the KAP was administered for the first time to students in Kansas. The KAP operational assessments currently include Mathematics and ELA in Grades 3 – 8 and 10. Multiple versions (*forms*) of the KAP were used in each subject and grade level. Several changes are expected in future KAP assessments. Some of the changes expected in 2016 for the KAP assessments include:

- Listening items (questions about a series of one- to two-minute auditory stimuli consisting of a person speaking or of a conversation) will be added in ELA.
- Writing performance tasks will be added. Prompts will follow a student's reading and analysis of complex, grade-level texts that may be Narrative, Informative/Explanatory, or Opinion/Argumentative.
- A multistage adaptive testing format.

Because of the large number of tables and figures included in this technical manual, it was created using the rmarkdown package[2] with the Tufte-style template[3]. These packages integrate data analysis results generated by the statistical program R[4] with text typeset with LaTeX. This also allows this document to be easily reproducible across years and assessment programs.

[2] http://cran.r-project.org/web/packages/rmarkdown/index.html

[3] https://tufte-latex.github.io/tufte-latex/

[4] R Core Team (2016). URL http://www.R-project.org/

Part I

Statewide System of Standards and Assessments

1
Introduction

1.1 Purposes of KAP

The four main purposes of the KAP are to:

- Measure specific claims related to the *Kansas College and Career Ready Standards* (KCCRS) as identified in the Performance Level Descriptors (PLDs);
- Provide information for calculating Annual Measurable Objectives (AMOs) and for state accreditation;
- Report individual student scores along with the student's performance level; and
- Provide subscale and total scores that can be used with local assessment scores to assist in improving a building or district's programs in the tested content areas.

Additional details about the KCCRS are provided in an upcoming chapter.

1.2 Tests in KAP

The following content areas have KCCRS standards and are, or will be in the near future, part of the KAP:

- Mathematics in Grades 3 – 8 and 10
- ELA in Grades 3 – 8 and 10, including:
  - Reading
  - Writing
  - Listening
- Science in Grades 5, 8, and 11
- History, government, and social studies (HGSS) in Grades 6, 8, and 11

Note that the new Science and HGSS standards will be assessed starting in the spring of 2016. Listening will also be integrated into the ELA tests at that time. This technical manual focuses on the new tests in mathematics and ELA, as those assessments were operational in 2015.

1.3 Statutory Authority

1.3.1 State Accountability

The state statutory authority behind KAP is Kansas 2014 Statute Chapter 72, Article 63, §39. Primary elements in this statute are for the state board of education to:

- Design and adopt a school-performance accreditation system based upon improvement in performance that reflects high academic standards and is measurable;
- Establish curriculum standards that reflect high academic standards for the core academic areas of mathematics, science, reading, writing, and social studies; and
- Provide statewide assessments in the core academic areas and determine performance levels on the statewide assessments.

It is further noted that the performance levels should represent high academic standards in the academic area at the grade level to which the assessment applies.

For additional information see www.kslegislature.org/li_2014/b2013_14/statute/072_000_0000_chapter/072_064_0000_article/072_064_0039_section/072_064_0039_k/

1.3.2 Federal Accountability

At the federal level, it is the *Elementary and Secondary Education Act* (ESEA) that supports Kansas's efforts to establish challenging standards, develop aligned assessments, and build accountability systems for districts and schools that are based on educational results.

1.4 Inclusion

Kansas is committed to including all students in KAP. Some notable exceptions include:

- Students serving long-term suspension
- Students who were truant for more than two consecutive weeks at the time of testing
- Students who experienced catastrophic illnesses or accidents
- Students who moved during testing

- Students who were incarcerated

In addition to these exceptions, English language learners (ELLs) who are recent arrivals to the United States are required to take the KAP mathematics tests, but their results count only toward participation. Further, ELLs are required to take the Kansas English Language Proficiency Assessment–Placement (KELPA–P) in lieu of the KAP ELA.

One percent (1%) of Kansas students take the Dynamic Learning Maps (DLM) alternate assessment. Other special-needs students with IEPs, 504 plans, or SIT plans take the KAP test but can have accommodation(s) consistent with their personal-needs profile (PNP). If an accommodation was given that is not allowed (e.g., reading to student on the KAP ELA test) the student was treated as *not tested*.

1.5 Participation Data

Detailed participation statistics by various subgroups are provided elsewhere in this technical manual. However, KAP participation rates were excellent, overall. Of the nearly 245,000 students in Grades 3 – 8 and 10, only 1,360 were not tested.

See the chapter on demographics, inclusion, and participation for more information.

1.6 FERPA

The Family Educational Rights and Privacy Act (FERPA) affords parents and students certain privacy rights with respect to students' educational records. These rights extend to students in all grade levels, from preschool through postsecondary education. FERPA applies to all schools that receive funds under any applicable program of the U.S. Department of Education.

Staff who work on the KAP and require access to student data take seriously the protection of student privacy and confidentiality. KAP data records are maintained on a secure server. Access to KAP data is limited to staff with specific responsibilities. Staff with data access undergo yearly FERPA training. FERPA compliance officers maintain records related to staff training and certification.

Part II

Academic Achievement Standards and Reporting

2
Academic Content Standards

Kansas's coherent and rigorous academic content standards are known as the *Kansas College and Career Ready Standards* (KCCRS). An overview of KCCRS development follows, describing the process and timeline for the development of content standards, the connection of Kansas standards to national standards, participant information, and the approval of the governing authority.

The subject area KCCRS may be viewed here: http:/www.ksassessments.org/about-kap

2.1 Development of Content Standards

Standards in Kansas were developed by a committee of Kansas educators. These standards help schools prepare students, providing the knowledge and skills needed to pursue higher education or better careers and to compete in an increasingly competitive and global work environment.

2.2 Process and Timeline for Development

Under the direction of and feedback from Kansas educators, the KCCRS were adapted from the *Common Core State Standards* (CCSS). Beginning in November 2009, KSDE received drafts of the CCSS and provided feedback to the Council of Chief State School Officers (CCSSO). From January 2010 to August 2010, Kansas educators who served on the ELA CCRS Committee provided feedback to the CCSSO and other groups involved in the development process; this feedback was incorporated into subsequent drafts of the CCSS.

See http://www.corestandards.org/wp-content/uploads/Math_Standards1.pdf and http://www.corestandards.org/wp-content/uploads/ELA_Standards1.pdf.

2.3 Convergence and Divergence with National Standards

THE CCSS:

> define what students should understand and be able to do by the end of each grade. They correspond to the College and Career Readiness (CCR) Anchor Standards [in the KCCRS] by number. The CCR and grade-specific Standards are necessary complements—the former providing broad standards, the latter providing additional specificity—that together define the skills and understandings that all students must demonstrate.[1]

[1] *Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects*, 2010, p. 10

The main difference between the CCSS and the KCCRS are the *Kansas 15%*; the purpose of which is to emphasize concepts and teaching philosophies that are important in Kansas. Although most of the concepts included within these Standards are mentioned in the CCSS, KSDE wanted to highlight the importance of each one by including it in the KCCRS.

KSDE added the Anchor Standards for Literary Learning, as well as four other Anchor Standards (two in Reading and two in Writing), as part of the Kansas 15%. As outlined in the CCSS document, the KCCRS are

> divided into strands. K – 5 and 6 – 12 ELA have Reading, Writing, Speaking and Listening, and Language strands; the 6 – 12 history/social studies, science, and technical subjects section focuses on Reading and Writing. Each strand is headed by a strand-specific set of College and Career Readiness Anchor Standards that is identical across all grades and content areas.[2]

[2] *Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects*, 2010, p. 8

For math, the Kansas additions to the CCSS were for Probability and Statistics and Algebraic Patterning. These two topics were *set aside from the detail of the main document*[3] so that each school and/or district could decide how to incorporate each topic.

[3] *KCCRS*, 2010, p. 8

The KCCRS are written for instructional use in the classroom, and teachers are expected to implement the breadth of the Standards. The Standards were reorganized for assessment purposes. Currently, the KAP does not include items measuring any part of the Kansas 15%.

2.4 Participants

COMMITTEE MEMBERS involved in development of the Kansas additions to the *Common Core State Standards for Mathematics* include:

- Jerry Braun, USD 489
- Pat Foster, USD 341
- Melisa Hancock, Kansas State University
- Marjorie Hill, University of Kansas

- Fred Hollingshead, USD 450
- Laura Ortiz, USD 457
- Allen Sylvester, USD 501
- Debbie Sylvester, USD 320
- Debbie Thompson, USD 259

ELA committee members include:

- James Heimann, USD 500
- Carolyn Boyd, USD 469
- Sandee Morris, USD 336
- Andy Anderson, Johnson County Community College
- Dianne Seltzer, KCMO
- Karla Reed, USD 231
- Vicki Seeger, USD 345
- Linda Ziegler, USD 293
- Judy Beemer, USD 475
- Bev Nye, USD 253
- Glynn Bennion, USD 466
- Deb Larson, USD 450
- Rebecca Hochstien, USD 265
- Sheryl Plattner, USD 441
- Stephanie Barnhill, USD 230
- Julie Dick, USD 375
- Leigh Ann Roderic, USD 457
- Kristi Orcutt, ESSDACK
- Ruthann Harris, USD 259
- Julie Aikins, USD 413
- Nancy Kent, USD 512
- Trinity Davis, Pittsburg State University

2.5 Approval by Governing Authority

In September 2010, the Standards were presented to the Kansas State Board of Education. On October 10, 2010, the Board adopted the KCCRS for use in Kansas.

3
Standard Setting

The Achievement and Assessment Institute (AAI) conducted standard setting for the KAP using the Bookmark Method during a workshop in Topeka on July 21—24, 2015. The main goals of that event were to establish the cut scores that differentiate:

- Level–1 performance from Level–2 performance;
- Level–2 performance from Level–3 performance; and
- Level–3 performance from Level–4 performance.

These three cut–scores establish the four performance levels for the assessment.

3.1 Overview of Bookmark Method

The Bookmark Method[1] is widely used in K—12 educational assessment contexts. Items are arranged from easiest to hardest based on empirical IRT item-parameter estimates and are placed in an ordered–item booklet (OIB).Panelists review items in order and place a bookmark at the page in the OIB where they believe the *just-barely* examinee at a given level would not have a particular probability of answering the item correctly.

[1] Mitzel, Lewis, Patz, & Green, 2001

The Bookmark Method uses IRT scaling to place items and students on the same scale. When the assumptions of IRT hold, a student's test score can be used to provide the probability that the student will answer a given multiple-choice item correctly, or, in the case of polytomously scored items, obtain a given score point.

The response probability (RP) value used in this event was 0.67. As an example of how this RP was applied, panelists may have been asked the following question: Would at least 20 out of 30 just-barely Level 3 students be able to answer this item correctly?

3.1.1 Rationale for Using the Bookmark Method

The KSDE and AAI considered the application of the Bookmark Method to be appropriate for two primary reasons. First, the KAP item pool was sufficiently large and covered a broad ability range. Second, the IRT item-parameter estimates had good precision because many students had taken the tests. Using assessment data collected during an operational administration, items were ranked in difficulty. Items were selected to eliminate noteworthy gaps in item difficulty.

3.1.2 Calculating Cut Scores

To determine the cut scores for the different performance levels, the median cut scores for level and across all panelists were calculated and then converted to an ability (theta) value that represented the group's estimated theta cut score for each performance level. Medians are often preferred over means in this context because they reduce the influence of any extreme judgments.

3.2 Panelist Demographics

Appropriate selection and training of panelists is crucial to the success of any standard-setting event. Considering several aspects of panel diversity (e.g., ethnicity, gender, geographic area, teaching experience, and role), KSDE recruited 117 educators to be panelists for the standard-setting event.

For the panelist nomination process, KSDE worked with constituent groups to nominate qualified teachers for the standard–setting event. The criteria for nominated teachers included: 1) availability for the full four-day event; 2) willingness to volunteer for the event; 3) teaching experience in the content area of nomination (e.g., math or ELA); 4) teaching experience in the grade level of nomination (e.g., Grades 3—8 or high school); 5) knowledge of the KCCRs; and 6) experience with ELL or special–education students. At the high school level, KSDE recruited teachers who had taught dual–enrollment college courses, AP/IB courses, and adjunct courses at the high–school level or in higher education.

For the panelist selection process, KSDE identified several preferred criteria. Selected panelists should represent:

- all 10 districts
- priority and focus schools
- a cross-section of the state's districts with regard to size, setting, and socioeconomic conditions
- a range of teaching experience

KSDE also gave first preference to teachers who did not participate in item reviews or serve on the PLD committee. Other factors considered in panelist selection included current licensure type, content endorsements, ELL or special education endorsements, gender, race or ethnicity, and teaching experience. At minimum, each panel had two tables composed of six panelists per table.

The full KAP Standard–Setting Report includes information about panelists' demographic composition and experience.

3.3 Performance–Level Descriptors (PLDs)

The PLDs provided critical information for panelists to consider when they made their bookmark judgments. KSDE staff and about 40 Kansas educators drafted the grade-specific PLDs for all four levels during October of 2014.

AAI staff with in-depth knowledge of the KCCRS first developed a written statement about student academic performance (not the student). Staff adhered to the cognitive alignment of the standards (depth of knowledge, cognitive complexity, levels of thinking, scope of skills, inquiry vs. process, etc.). All of the standards describe what all students should know and be able to do regardless of the specific items that actually appear on the test each year. In the development and review of PLDs, there was no discussion of just–barely students. The focus was on what all students should know and be able to do. The PLDs were submitted to the KSDE for adoption along with the recommended cut scores.

PLDs for the four performance levels (Level 1, Level 2, Level 3, and Level 4) are provided on the KAP Assessments website. Note that the number of students in Levels 3 and 4 are important for accountability.

For ELA
http://ksassessments.org/languagearts.
For math
http://ksassessments.org/math.

3.4 Procedural Overview

A comprehensive description of the standard–setting procedures is included in the full KAP Standard–Setting Report. Panelists were thoroughly trained before engaging in the three standard–setting rounds. Before placing bookmarks, panelists:

- took the operational test items;
- defined the just-barely student at each performance level;
- engaged in a practice activity;
- described the knowledge and skills required to answer each test item; and

- completed a form to indicate their readiness for the standard–setting activities.

Panelists completed three rounds of bookmark placements. For each round, panelists placed bookmarks for Level 3, then for Level 4, then for Level 2. After Round 1, panelists reviewed their results and then discussed table-level results. After Round 2, panelists reviewed table-level results, room-level results, and impact data. After the final round, panelists again reviewed room-level results and impact data.

Panelists' Evaluation Form results are discussed in the validity chapter.

3.5 Policy Review

AAI STAFF PRESENTED RESULTS from Round 3 to a policy review panel, the objective of which was to ensure the reasonableness of cut scores across grades. Approximately 40 educators participated in this phase. The deputy commissioner individuals for the policy review panel. Additionally, KSDE selected several teachers from the standard–setting panels to attend the policy review meeting.

At the beginning of the meeting, the policy review facilitator reviewed the Bookmark Method procedures with the policy group. The facilitator provided the assertions and context for the standard–setting meeting, reviewed information about the PLDs, provided an overview of the steps in the standard–setting process, and discussed the materials that panelists used.

The facilitator then presented impact data from the Round 3 bookmark placement results from ELA and math. Policy review panelists then provided their feedback about the process and the impact data.

The facilitator then presented the smoothed and unsmoothed results (cuts and impact data) for all grades. Panelists considered and discussed the results and then recommended reasonable changes.

3.6 Approval by the Governing Authority

The Kansas State Board of Education approved the cut scores on September 8, 2015.

3.7 Final Cut–Scores

THE TABLES DISPLAY final cut–scores and performance–level percentages for math and ELA.

Grade	Level 1/2 Cut	Level 2/3 Cut	Level 3/4 Cut
3	276	300	329
4	266	300	331
5	273	300	326
6	273	300	329
7	266	300	342
8	274	300	336
10	275	300	333

Table 3.1: Math SS Cuts

Grade	Level 1/2 Cut	Level 2/3 Cut	Level 3/4 Cut
3	276	300	327
4	271	300	335
5	275	300	326
6	277	300	336
7	275	300	335
8	265	300	334
10	269	300	334

Table 3.2: ELA SS Cuts

Grade	Level 1	Level 2	Level 3	Level 4	Level 3 + 4
3	0.122	0.353	0.366	0.160	0.525
4	0.135	0.502	0.280	0.083	0.363
5	0.231	0.426	0.238	0.105	0.342
6	0.203	0.462	0.249	0.086	0.335
7	0.150	0.551	0.266	0.034	0.299
8	0.363	0.401	0.194	0.042	0.237
10	0.372	0.381	0.199	0.048	0.247

Table 3.3: Proportion of Students in each Performance Level by Grade for Math

Grade	Level 1	Level 2	Level 3	Level 4	Level 3 + 4
3	0.196	0.327	0.345	0.132	0.477
4	0.105	0.333	0.451	0.112	0.563
5	0.176	0.328	0.347	0.149	0.496
6	0.265	0.325	0.372	0.038	0.410
7	0.251	0.344	0.376	0.029	0.405
8	0.203	0.493	0.281	0.023	0.304
10	0.240	0.442	0.296	0.022	0.318

Table 3.4: Proportion of Students in each Performance Level by Grade for ELA

4
Item and Test Scores

This chapter provides information about KAP items and test scores, including scaled scores, performance levels, and claim scores, as well as appropriate and inappropriate uses of these scores.

4.1 KAP Items

4.1.1 Item Types

The KAP item types are listed below. As shown in the tables on the right, the multiple-choice keyed (MC-K) items are the most frequently used item type by roughly a two-to-one margin in most grades. This is not surprising because the MC-K item format is well established and known to be efficient for measuring a broad range of content. In other words, more material can be sampled with the MC-K format than almost any other item format available. These items can also be scored rapidly with few processing errors. Innovative task package (ITP) items were the second most common item format. The least used item format was constructed-response (CR) in math and multiple-choice multiple-select (MC-MS) in ELA. Although there were no CR items in ELA this year, next year's assessment will include extended constructed-response (ECR) items in the form of multidisciplinary performance task (MDPT) writing prompts.

- selected-response (SR) items
  - multiple-choice keyed (MC-K) items: *selection items with one correct response*
  - multiple-choice multiple-select (MC-MS) items: *selection items with multiple correct responses possible*
- constructed-response (CR) items

Table 4.1: Item Counts by Item Type and Grade for Math

Grade	CR	ITP	MC-K	MC-MS
3	19	37	144	20
4	10	58	141	29
5	15	40	141	22
6	11	33	130	13
7	13	31	135	19
8	9	32	137	16
10	2	21	148	18

Table 4.2: Item Counts by Item Type and Grade for ELA

Grade	ITP	MC-K	MC-MS
3	78	209	24
4	65	194	29
5	51	203	35
6	60	198	27
7	66	173	30
8	52	177	21
10	73	188	30

  - short constructed-response (SCR) items: *short free-response items typically worth one point*
  - extended constructed-response (ECR) items: *long free-response items typically worth more than one point*

ECR items were not used in 2015 but will be included in the future.

- innovative task package (ITP) items
  - background graphic items
  - drop-down menu items
  - labeling items
  - matching-line items
  - matrix interaction items
  - multiple-drop bucket items
  - ordering items
  - partition items
  - select-text items
  - plotting point items
  - straight line items
  - Venn diagram items

ITP items are also known as technology-enhanced items (TEIs) The last three ITP task types are currently used in math only. Examples of the ITPs used on the KAP assessment may be viewed through the KAP practice tests. Instructions are provided at: www.ksassessments.org/practice-tests.

4.1.2 Item Scores

As seen in the tables on the right, math item scores range from zero to two, and ELA item scores range from zero to three. One-point items were used most frequently, which is not surprising given that the MC-K items—worth one point in value—were the most frequently used item format. There were a handful of two-point items in math and three-point items in ELA. Other item types are not always associated with specific score values; content is the primary consideration in assigning point values to non-MC-K items.

Table 4.3: Item Counts by Item Points and Grade for Math

Grade	1-Point	2-Point
3	219	1
4	233	5
5	216	2
6	187	0
7	198	0
8	192	2
10	188	1

Table 4.4: Item Counts by Item Points and Grade for ELA

Grade	1-Point	2-Point	3-Point
3	278	33	0
4	243	45	0
5	237	52	0
6	248	37	0
7	211	50	8
8	210	39	1
10	236	48	7

4.1.2.1 Decimal Item Scores

Maximum item scores do not always indicate the number of score categories because decimal scores are possible in the KAP. An item with a maximum score of one point might have decimal score values of 0.00, 0.33, 0.67, and 1.0. Assignment of specific decimal score values for any given item is based on content considerations, but for many items the number of possible score values is equal to the number of response categories for the item, plus one (to account for a score of 0).

A two-point item with four response categories might have five $(4 + 1 = 5)$ decimal score values: 0.0, 0.5, 1.0, 1.5, and 2.0.

Item counts by the number of response categories and grade are provided below. Most items have two response categories because most items are MC-K with dichotomous scores (0, 1). In math, most items have less than six response categories. In ELA most items have less than five response categories.

Grade by RC	2	3	4	5	6	7	8	9	11
3	191	8	14	5	0	1	0	1	0
4	177	11	25	13	7	4	0	0	1
5	174	13	15	14	0	2	0	0	0
6	155	7	17	8	0	0	0	0	0
7	168	11	10	5	3	0	1	0	0
8	164	7	9	11	1	2	0	0	0
10	163	9	14	3	0	0	0	0	0

Table 4.5: Item Counts by Response Category (RC) and Grade for Math

Grade by RC	2	3	4	5	6	7	8	9
3	269	30	7	1	1	2	1	0
4	240	30	17	1	0	0	0	0
5	235	45	5	3	0	0	1	0
6	235	39	7	2	1	1	0	0
7	212	35	18	2	1	0	0	1
8	213	23	7	5	0	0	1	1
10	239	32	15	2	0	3	0	0

Table 4.6: Item Counts by Response Category (RC) and Grade for ELA

4.2 Total Test Scores

The KAP uses two different types of test scores—scaled scores and performance levels—that have different properties based on their specific purposes and uses.

4.2.1 Raw Scores

Raw scores (RSs) do not appear on student score reports. However, they are used by test developers behind the scenes and are important intermediate scores, which are transformed into scaled scores. An RS is the sum of points a student earns over the operational test items. Although decimal scores are possible on items, the final RSs are rounded to the nearest integer.

When creating the integer scores, decimals ending in exactly 0.5 are rounded up.

RSs have limited use in isolation. For example, RSs have to be referenced against the total number of possible points on the test (e.g., an RS of 15 on a 20-point test is different than an RS of 15 on a 30-point test). In addition, RSs depend on the overall difficulty of the test items (e.g., an RS of 15 on a test with 20 easy items is different than an RS of 15 on a test with 20 difficult items). For these reasons, RSs are not used on KAP score reports.

4.2.2 Scaled Scores

A scaled score (SS) is a transformed RS. The specifics of the transformation processes for the KAP are discussed in the chapter on scaling. When students take the same test form, the relationship between the RSs and SSs is *monotonically increasing*.

Monotonically increasing means that as RSs increase, any SS must be greater than or equal to the SS that precedes it.

Using SSs instead of RSs produces more general, interpretable, and equitable results. Specifically, the SSs remove the effects of test length and item difficulty.

The transformation of RSs to percent-correct scores removes the effect of test length, but does not adjust for item difficulty differences.

Some believe that SSs lend themselves to *interval-level* interpretations. A true interval-level SS would mean that a difference of five SS points represents the same magnitude of difference regardless of where the SSs are on the scale (e.g., from 220 to 225, 300 to 305, or 375 to 380).

Not everyone believes that scaled scores possess interval-level properties. There is much greater consensus that RSs don't support interval-level interpretations.

When test SSs are properly linked across years, any given SS value (e.g., 300) for a particular grade-level and subject-area test (like Grade 4 ELA) should have the same meaning in the current year as it had in previous years. An increase in the median scaled score for Grade 4 ELA from last year to the current year means that student performance improved. In contrast, an increase in median RSs might simply meant that overall test difficulty decreased.

The given example is not an endorsement of conducting a trend analysis with only two years of data. Further, small differences may not be statistically or practically significant.

For more information about KAP scaled scores, see (1) the scaling chapter, which provides information on the development of the KAP scaled-score system, including transformation formulas, rounding rules, and general scale characteristics (e.g., minimum and maximum possible values) and (2) the operational statistics chapter, which provides descriptive statistics for scaled scores.

4.2.3 Performance Levels

KAP individual score reports also include the student's performance level, of which four classifications are possible (Level 1, Level 2, Level 3, and Level 4). The cut scores on the scaled-score metric (i.e., the lowest possible scaled scores to enter Level 2, Level 3, and Level 4) are presented elsewhere in this manual, but are repeated here for convenience. The operational statistics chapter provides the percentage of students that fell in each performance level.

The KAP cut scores are located online at: http://www.ksassessments.org/cutscores

Grade	Level 1/2 Cut	Level 2/3 Cut	Level 3/4 Cut
3	276	300	329
4	266	300	331
5	273	300	326
6	273	300	329
7	266	300	342
8	274	300	336
10	275	300	333

Table 4.7: Math SS Cut Scores

Grade	Level 1/2 Cut	Level 2/3 Cut	Level 3/4 Cut
3	276	300	327
4	271	300	335
5	275	300	326
6	277	300	336
7	275	300	335
8	265	300	334
10	269	300	334

Table 4.8: ELA SS Cut Scores

4.2.4 Performance-Level Descriptors

The KAP performance-level descriptors (PLDs) are the primary way to attach meaning to test results. For each of the quantitative ranges of scaled scores that create the four performance levels, the PLDs provide verbal, qualitative descriptions of what students know and can do. These qualitative descriptions of academic skills are emphasized to students, parents, and teachers as the primary mechanism for interpreting student scores.

The KAP PLDs are located online at: http://www.ksassessments.org/pld

4.3 Claim Scores

A CLAIM SCORE DESCRIBES the performance of a student, school, or district on a particular *claim* (a related set of standards from the *Kansas College and Career Ready Standards*). For the KAP, claim scores are reported as scaled scores. Scaling procedures for claim scores are overviewed in the scaling chapter.

Claim scores should not be compared across years because there is not a strong statistical link at this level. Different claim scaled scores can be compared within the same test; however, the low reliability of many of the claim scores will limit such comparisons for individual students. Claim scores can be helpful in identifying group-level

The reliability chapter provides more information about the precision of the claim scores.

strengths and weaknesses because the reliability of aggregated claim scores is usually better than that of individual claim scores. Claim scores can suggest group strengths or weaknesses relative to another reference group. This can be done by comparing the median claim score of a school against the median claim score of another reference group (e.g., the state or district). A school (or district) may also compare pairs of claim scores. However, error bands should be considered when doing this.

Test-score error bands are discussed in the reliability chapter.

4.4 Score Uses

4.4.1 Some Appropriate Score Uses

Generalizations from the KAP test results may be made about the specific content domains measured by the KAP, as outlined in the *Kansas College and Career Ready Standards* (KCCRS). However, any instructional and program interventions at the student or group level should be based on as much information from other sources as possible. Multiple information sources will provide a more complete picture of performance and avoid the threat of mono-operational bias in program evaluation.

4.4.1.1 Individual Students

Scaled scores on the KAP indicate a student's achievement of the KCCRS. Scaled scores are primarily used to determine student performance-level classifications (a *criterion-referenced inference*). Scaled scores can also be used to compare the performance of an individual student to the aggregate performance of a school or district. When comparing the performance of individual students to these reference points, test-score standard errors should be considered because scaled scores are estimates of students' achievement that contain random error.

4.4.1.2 Groups of Students

Test results can be used to evaluate performance over time. Average (median or mean) scaled scores can be compared across administrations within the same grade and subject area to indicate whether student performance is improving across years. Generally, such trend analyses benefit from using average results from as many test administration years as possible. Different cohorts of students are used (i.e., the same student or students are *not* tracked across grade levels). All scores can be analyzed within the same subject and grade from any single administration to determine which demographic or

program group had, for example, the highest average performance or the highest percentage of students at or above the Level 3 standard. Average group claim scores can help evaluate academic areas for relative strengths or weaknesses. These scores can help identify areas that warrant further evaluation.

4.4.2 Some Cautions for Score Use

4.4.2.1 Extreme Scores

To minimize confusion and potential misinterpretation, the minimum and maximum scaled scores possible on the KAP tests have been fixed so they do not change between administrations. Student scores near the minimum or maximum ends of the score range will have large standard errors of measurement and, therefore, such scores should be viewed cautiously. The minimum and maximum scaled scores only provide rough estimates of a student's ability. For instance, if the maximum score for the KAP Grade 6 mathematics test was 380 and a student achieved this score, it could not be determined whether the student could have achieved an even higher scaled score.

4.4.2.2 Total-Test Scaled Scores from Different Tests

KAP scaled scores appear to be the same across tests, but they are not. Scaling was conducted for each grade-level and subject-area test separately. KAP scaled scores are not status indicators in the same sense as percentile ranks (or scales that are essentially transformations of percentile ranks) and, therefore, cannot be used to profile relative strengths and weaknesses across subject areas. For example, a student with scaled scores of 320 in Grade 4 math and 310 in Grade 4 ELA is not necessarily better in mathematics. The KAP scaled scores are not a developmental or vertical scale. This means that comparisons of student growth across grades are not appropriate. For example, a 330 in Grade 4 ELA and a 330 in Grade 5 ELA do not indicate that a student had no achievement growth in ELA from Grade 4 to Grade 5.

4.4.2.3 Claim Score Caveats

While in rare cases claim scores might facilitate comparisons of student strengths and weaknesses across claims, several factors merit caution. Many claim scores are not reliable enough for such comparisons and the scaling procedures that put the claim scores on the same scale do not correct this issue.

A comparison between claim scores and total test scores may produce results that seem unusual. For example, it is possible for a student to score slightly below the state median on all claim scores but

score slightly above the state median on the total test. The reverse scenario is also possible. This seemingly odd circumstance can happen, in part, because the claim scores are correlated, meaning the distributional properties of the total score depend not only on the variances of the claim scores, but also on the covariances among the claim scores. The fact that the total scores and claim scores were scaled separately, and then rounded before final reporting, can contribute to this oddity as well.

The covariance explanation alone may be too technical for some. A rough analogy about track and field competetions may aid one's understanding of the covariance explanation. A school track team can place first overall in a track meet although no team member placed first in any of the individual events.

4.4.2.4 Using KAP Results for Other Purposes

Should KAP results be used for placement decisions, such as eligibility for gifted-talented programs or other similar services? Consider questions such as these:

- What is the maximum observed KAP scaled score?
- What KAP score represents the 90th percentile?

Perhaps the motivation behind questions like these concerns special program eligibility at certain schools.

Other uses or inferences based on KAP results may or may not be valid, and the evidence and arguments provided in the validity chapter may not adequately address such use cases or interpretations. According to Standard 1.4 from the *Standards for Educational and Psychological Testing*[1], if a test is used in a way that has not been validated, the user should justify the new use and collect evidence to support that use. Finally, a universal caveat for any test results is that they should not be the only source of information for placement and educational planning. Instead, other information about the student (e.g., other test-performance data) should be considered.

[1] AERA, APA, and NCME, 1999, p. 24

5
Score Reports

This chapter provides information about KAP score reports. Chapter 4 provides information about the scores themselves and inlcudes important caveats about score use. Readers should review Chapter 4 before continuing with Chapter 5. Images of actual KAP score reports appear in Chapter 5.

Personally identifying information has been redacted from the score reports shown in this chapter for the privacy of students, schools, and districts.

5.1 Reporting System

The following score reports are provided to students, parents, schools, and districts for the KAP mathematics and ELA tests:

- Individual Student Reports (ISRs)
- School Summary Reports (SSRs)
- School Detail Reports (SDRs)
- District Summary Reports (DSRs)
- District Detail Reports (DDRs)

Samples of these reports are provided below.

5.1.1 Subgroup Masking

When group n-counts are very small, individual identification is sometimes possible, even on roll-up summary reports. Various types of suppression logic are taken to prevent this identification. One type is to report a range of performance-level percentages instead of observed, actual percentages. For example, if only one student in a group of five students is in Level 4, the group's actual percentage is 20%. In a roll-up summary, however, the report gives a range of percentages instead (e.g., 0-40).

5.2 Individual Student Reports (ISRs)

ISRs are available for all students who took the KAP. The standard-setting event and KSDE's required approval of performance-level cut scores delayed the delivery of ISRs to educators from summary to early fall. Redacted pages from an ISR follow.

A student's total test performance is expressed by an image that resembles a meter. Both the total-test scaled score and the performance-level score are provided. A student's school and district median scaled scores, as well as the state median scaled score, are also given. A bar graph on page 2 of the ISR shows claim-score performance. School, district, and state medians are provided for reference.

To provide a measure of central tendency that is more robust to outliers, medians were used instead of means. Standard errors are provided for all scores. The student score SEM, simply denoted as *SE* on score reports, is the conditional standard error of measurement (CSEM) derived from the IRT scaling model. However, group standard errors are based on an estimate of the standard error for a median under normal distribution assumptions. The formula accounts for sampling error but not measurement error. The formula, which accounts for sampling error but not for measurement error, is equivalent to the well-known standard error for a mean, but multiplied by an extension factor of 1.253 to account for the additional sampling variability of the median.

See the reliability chapter for more information about CSEMs.

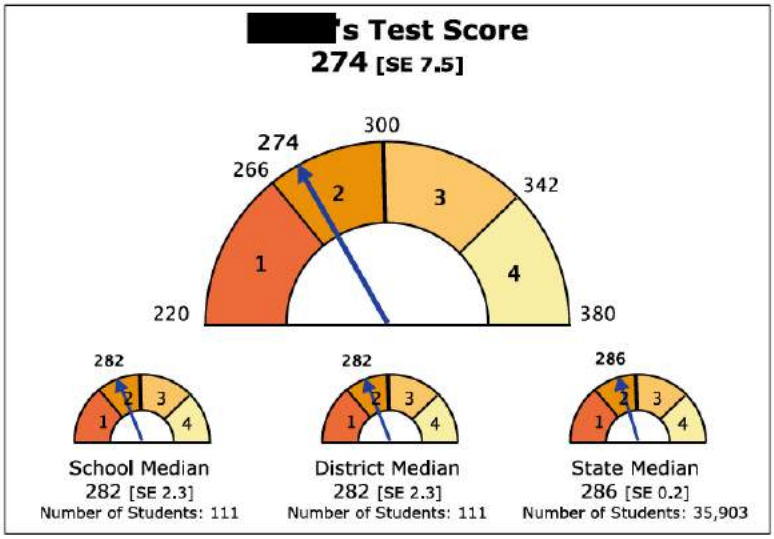
Figure 5.1: ISR Page 1

Student Report
Student: [redacted]
Student State ID: [redacted]
School Year: 2014–2015

KANSAS ASSESSMENT PROGRAM

Grade 7 Mathematics
School: [redacted]
District: [redacted]

This report has information about your student's Kansas Assessment Program (KAP) test scores. The KAP assessments measure a student's understanding of the Kansas College and Career Ready Standards at the student's grade level. The test contains questions that ask students to select the right answer as well as questions that ask the student to sort items, create graphs, or label pictures.



The first graph shows [redacted]'s overall score on the Mathematics test. The bands on the graph represent the four possible levels, with 4 being the highest level. The arrow shows [redacted]'s score.

The three smaller graphs show the performance of other seventh graders in [redacted]'s school, the school district, and the state. The median, or middle number in an ordered list of numbers, is used for these comparison graphs.

Performance Levels

Overall scores on the KAP test are divided into four performance levels. The levels range from 1 to 4, with 4 being the highest level. [redacted]'s score is in Level 2.

Level	Score Range	Level Name
4	342 - 380	Level 4
3	300 - 341	Level 3
2	266 - 299	Level 2
1	220 - 265	Level 1

The typical student who performs at this level can solve real-world problems involving proportional relationships presented in various formats; determine the constant of proportionality; convert between fractions and decimals; perform operations with rational numbers; factor and expand linear expressions with integer coefficients; add and subtract linear expressions with rational coefficients; and create and solve equations and inequalities with variables.

Explanation of Median and Standard Error

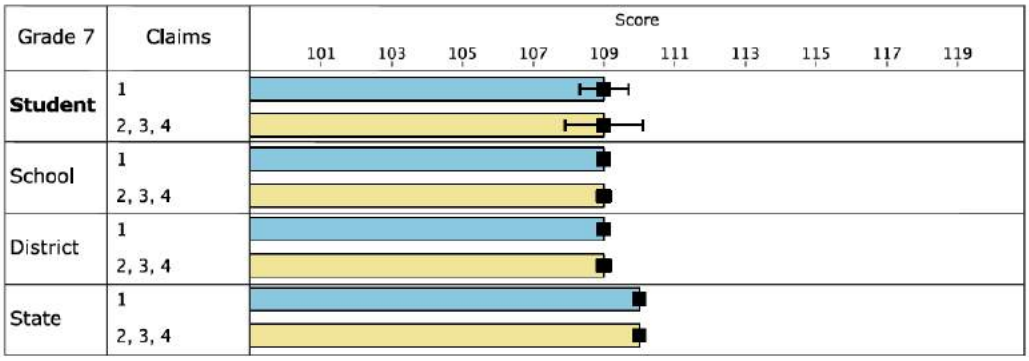
School, district, and state scores on this report are represented by the median score. A median is the middle number in an ordered list of numbers. For example, in the ordered list of scores {200, 210, 220, 230, 240, 250, 260}, the score of 230 is the median. The graphs show how the student's score compares to the median score for all students in the same grade who took the test in the school, district, and state.

Each score is also associated with a standard error of measurement (SE). The standard error around a student's score indicates how much a student's score might vary if the student took many equivalent versions of the test (a test with different items but covering the same content). The SE around the school, district, and state scores can be interpreted in a similar way. Standard error generally becomes smaller with larger comparison groups.

Figure 5.2: ISR Page 2

Student Report [redacted] Grade 7 Mathematics

Student's Relative Areas of Strength



This chart shows your student's performance relative to other students in the school, district and state on specific areas of the Grade 7 Mathematics test. Note that the scale is different from the overall test score. This information is not intended to be used to make instructional decisions because the number of items is too small. The bracket on either side of the bold score line represents the standard error.

Mathematics test questions cover four main areas (also called claims) of the Kansas Mathematics Standards. There are fewer questions on the test for Problem Solving, Communicating and Reasoning, and Modeling and Data Analysis. Therefore, these have been grouped together on the graph.

- Claim 1: Concepts and Procedures. These questions require students to explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.
- Claim 2: Problem Solving. These questions require students to solve a range of complex problems using knowledge, problem solving strategies, and mathematical tools.
- Claim 3: Communicating and Reasoning. These questions require students to explain their reasoning, defend their answers, critique the reasoning of others and ask clarifying questions.
- Claim 4: Modeling and Data Analysis. These questions require students to analyze complex, real-world situations and construct and use mathematical models to solve problems, as well as interpret their result in the context of a situation.

Additional Resources

For information on the Kansas College and Career Ready Standards, visit http://kap.cete.us/kccrs.
For information on the Kansas Assessment Program, visit http://ksassessments.org.
For the 2015 Interpretive Guide for score reports, visit http://kap.cete.us/ig.

5.3 School Summary Reports (SSRs)

SSRs are provided to schools and use bar graphs to provide summary information about the median total-test scaled score at each grade level in the school. District and state median scaled scores are provided as reference. The percentage of students in each of the four performance levels is shown by grade on page 2, with district and state results again provided as reference. Floating bar graphs present these results and clearly show the percentage of students at Levels 3 and 4, which is important for accountability purposes. Scanning the right edges of the bars reveals the groups with the highest percentage of students in these two levels. Redacted pages from an SSR follow.

Figure 5.3: SSR Page 1

School Summary Report
School: [redacted]
District: [redacted]

KANSAS ASSESSMENT PROGRAM

Mathematics
School Year: 2014–2015

This report has information about a school's scores from the Kansas Assessment Program. The tests measure students' understanding of Kansas College and Career Ready Standards at each grade using questions that ask students to select the right answer, sort items, create graphs, or label pictures. For sample test questions, see http://ksassessments.org/practice-tests.

Median School Performance

This chart compares the school's overall Mathematics test scores by grade to the median Mathematics test scores in the district and state.

Explanation of Median and Standard Error

School, district, and state scores on this report are represented by the median score. A median is the middle number in an ordered list of numbers. For example, in the ordered list of scores {200, 210, 220, 230, 240, 250, 260}, the score of 230 is the median. The graphs show how the student's score compares to the median score for all students in the same grade who took the test in the school, district, and state.

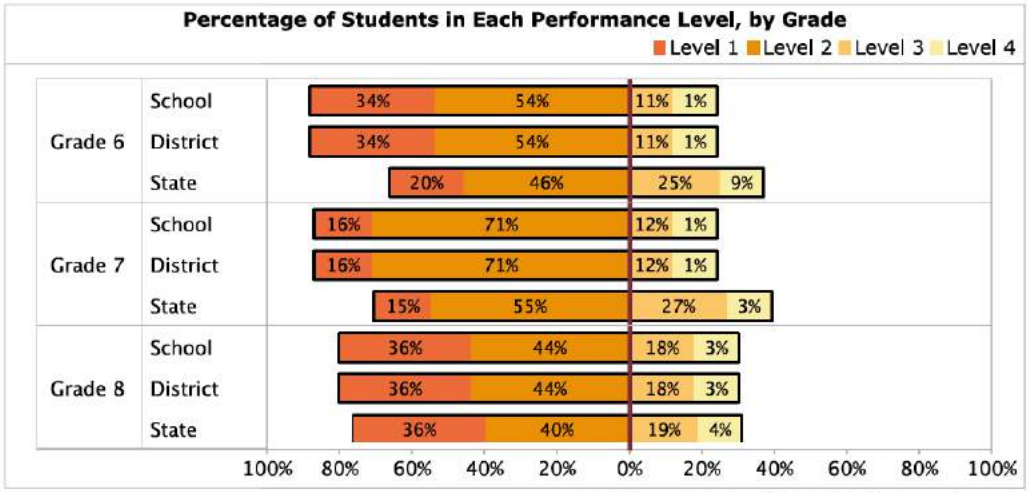
Each score is also associated with a standard error of measurement (SE). The standard error around a student's score indicates how much a student's score might vary if the student took many equivalent versions of the test (a test with different items but covering the same content). The SE around the school, district, and state scores can be interpreted in a similar way. Standard error generally becomes smaller with larger comparison groups.

Figure 5.4: SSR Page 2

School Summary Report [redacted] Mathematics

Overall scores on the KAP test are divided into four performance levels. The levels range from 1 to 4, with 4 being the highest level. Cut scores for levels 2 and 4 vary by grade.

The chart below compares the percentage of Mathematics students at the school by grade in each performance level to the district and state. Complete performance level descriptors with the cut scores can be found at http://ksassessments.org/pld.



*Percentages may not add to 100% because of rounding.

5.4 School Detail Reports (SDRs)

SDRs also are provided to schools. In contrast to SSRs, SDRs show school results by grade. Meter-style graphs illustrate total-test scaled score performance (school median with district- and state- referenced medians). Floating bar graphs illustrate the performance-level percentages. Claim-level results are provided using traditional bar graphs on page 2 of the report. District and state claim-score performances are provided as reference. Redacted pages from an SDR follow.

Figure 5.5: SDR Page 1

School Detail Report
School: [redacted]
District: [redacted]

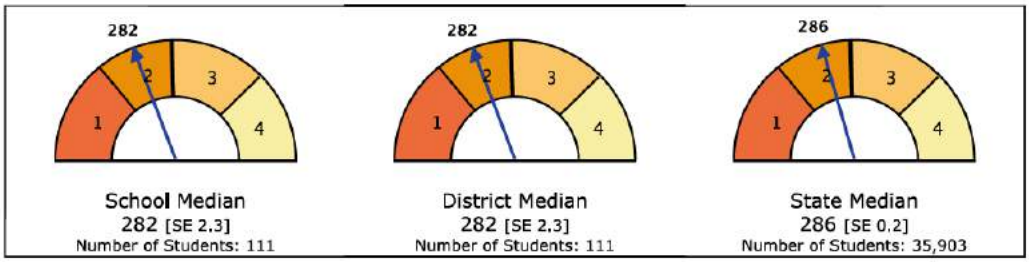
KANSAS ASSESSMENT PROGRAM

Grade 7 Mathematics
School Year: 2014–2015

This report has information about a school's scores from the Kansas Assessment Program. The tests measure students' understanding of Kansas College and Career Ready Standards at each grade using questions that ask students to select the right answer, sort items, create graphs, or label pictures. For sample test questions, see http://ksassessments.org/practice-tests.

School Median Score

The first graph shows the school's overall median score on the test, indicated by the arrow. The bands on the graph represent the four possible levels, with 4 being the highest level. The other graphs show the performance of seventh graders in the district and state. The median, or middle number in an ordered list of numbers, is used for these comparison graphs.

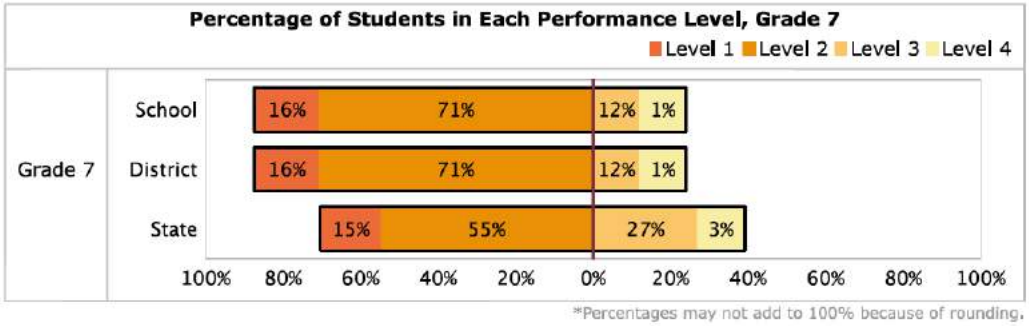


Performance Levels

Overall scores on the KAP test are divided into four performance levels. The levels range from 1 to 4, with 4 being the highest level. The school's median score is in Level 2.

Level	Score Range	Level Name
4	342 - 380	Level 4
3	300 - 341	Level 3
2	266 - 299	Level 2
1	220 - 265	Level 1

The following chart compares the percentage of seventh grade mathematics students in each performance level for school, district and state. Complete performance level descriptors can be found at http://ksassessments.org/pld.



*Percentages may not add to 100% because of rounding.

Figure 5.6: SDR Page 2

School Detail Report [redacted] Grade 7 Mathematics

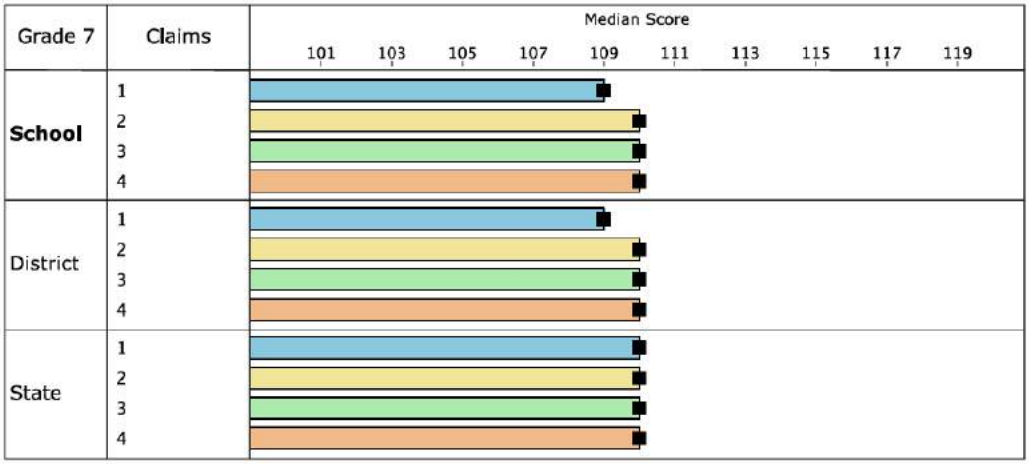
Explanation of Median and Standard Error

School, district, and state scores on this report are represented by the median score. A median is the middle number in an ordered list of numbers. For example, in the ordered list of scores {200, 210, 220, 230, 240, 250, 260}, the score of 230 is the median. The graphs show how the student's score compares to the median score for all students in the same grade who took the test in the school, district, and state.

Each score is also associated with a standard error of measurement (SE). The standard error around a student's score indicates how much a student's score might vary if the student took many equivalent versions of the test (a test with different items but covering the same content). The SE around the school, district, and state scores can be interpreted in a similar way. Standard error generally becomes smaller with larger comparison groups.

School Sub-Scores and Claims

This chart shows the school's performance on specific areas of the Grade 7 Mathematics test as well as the performance of the grade 7 students in the district and state. The bracket on either side of the bold score line represents the standard error, or how much a student's performance might vary if the student took many equivalent versions of the test.



Mathematics test questions cover four main areas (also called claims) of the Kansas Mathematics Standards.

- Claim 1: Concepts and Procedures. These questions require students to explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.
- Claim 2: Problem Solving. These questions require students to solve a range of complex problems using knowledge, problem solving strategies, and mathematical tools.
- Claim 3: Communicating and Reasoning. These questions require students to explain their reasoning, defend their answers, critique the reasoning of others and ask clarifying questions.
- Claim 4: Modeling and Data Analysis. These questions require students to analyze complex, real-world situations and construct and use mathematical models to solve problems, as well as interpret their result in the context of a situation.

Additional Resources

For the 2015 Interpretive Guide for score reports, visit http://kap.cete.us/ig.

5.5 District Summary Reports (DSRs)

DSRs are provided to districts and provide the same information as SSRs, but the only reference group provided is the state. Redacted pages from a DSR follow.

Figure 5.7: DSR Page 1

District Summary Report
District: [redacted]

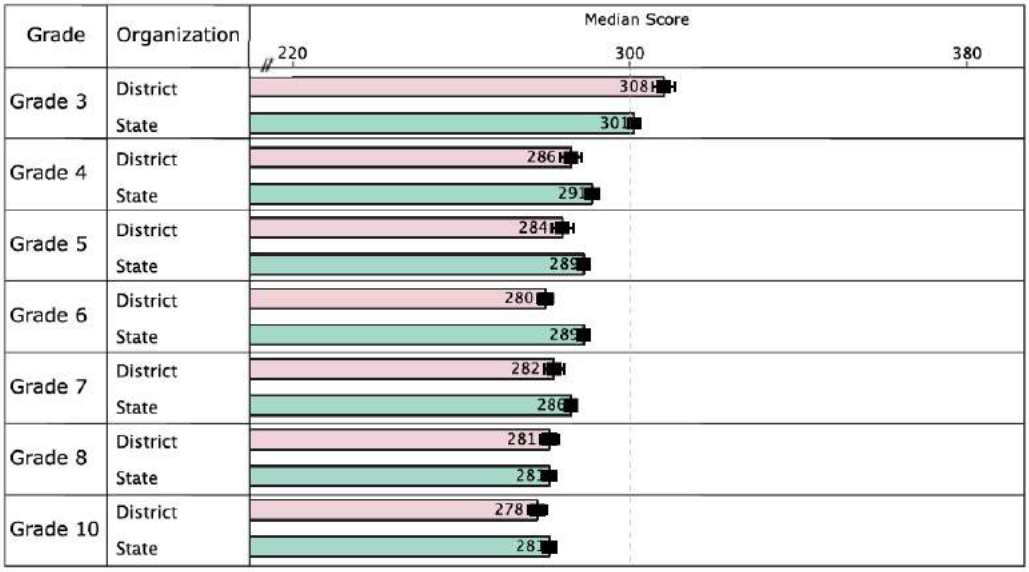
KANSAS ASSESSMENT PROGRAM

Mathematics
School Year: 2014–2015

This report has information about a district's scores from the Kansas Assessment Program. The tests measure students' understanding of Kansas College and Career Ready Standards at each grade using questions that ask students to select the right answer, sort items, create graphs, or label pictures. For sample test questions, see http://ksassessments.org/practice-tests.

Median District Performance

This chart compares the district's overall Mathematics test scores by grade to the median Mathematics test scores in the state.



Explanation of Median and Standard Error

School, district, and state scores on this report are represented by the median score. A median is the middle number in an ordered list of numbers. For example, in the ordered list of scores {200, 210, 220, 230, 240, 250, 260}, the score of 230 is the median. The graphs show how the student's score compares to the median score for all students in the same grade who took the test in the school, district, and state.

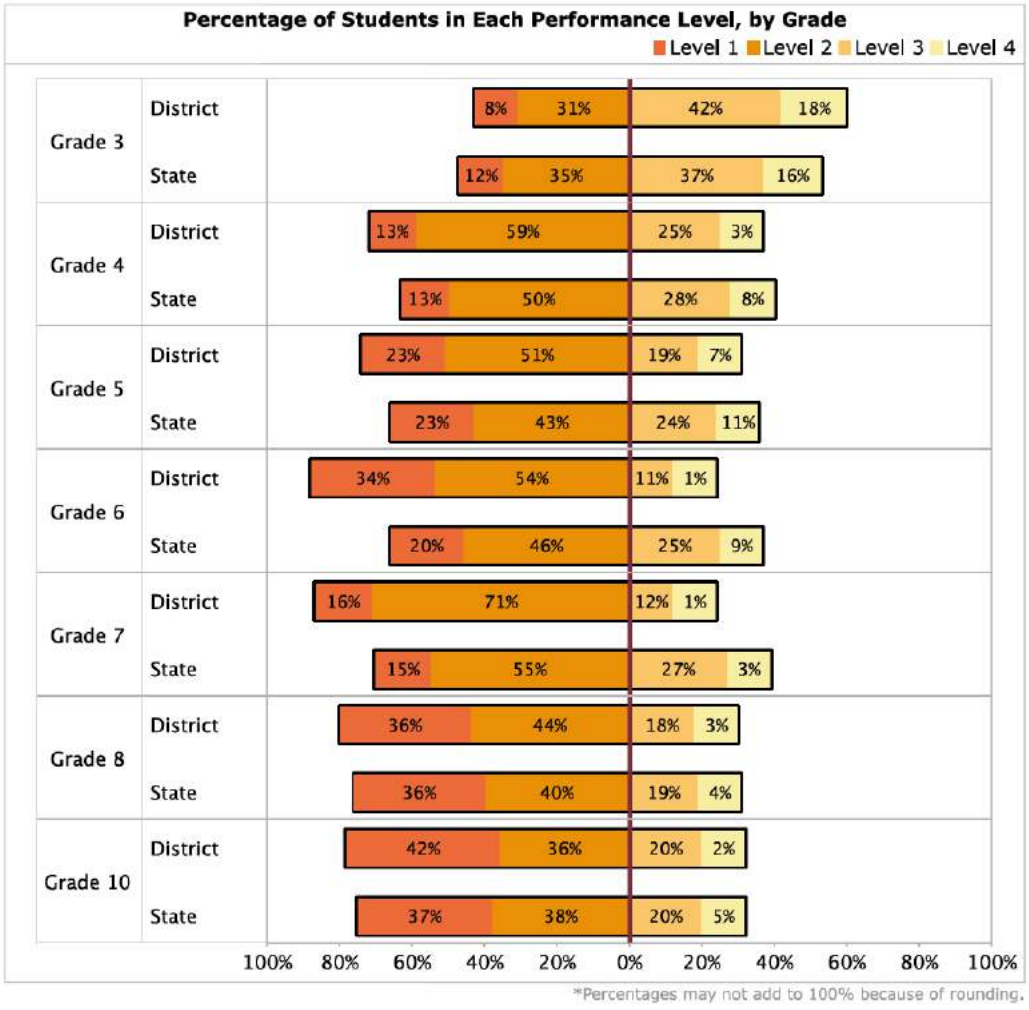
Each score is also associated with a standard error of measurement (SE). The standard error around a student's score indicates how much a student's score might vary if the student took many equivalent versions of the test (a test with different items but covering the same content). The SE around the school, district, and state scores can be interpreted in a similar way. Standard error generally becomes smaller with larger comparison groups.

Figure 5.8: DSR Page 2

District Summary Report [redacted] Mathematics

Overall scores on the KAP test are divided into four performance levels. The levels range from 1 to 4, with 4 being the highest level. Cut scores for levels 2 and 4 vary by grade.

The chart below compares the percentage of Mathematics students at the district by grade in each performance level to the state. Complete performance level descriptors with the cut scores can be found at http://ksassessments.org/pld.



*Percentages may not add to 100% because of rounding.

5.6 District Detail Reports (DDRs)

DDRs are provided to districts. DDRs are similar to SDRs, but the only reference group provided is the state. Redacted pages from a DDR follow.

Figure 5.9: DDR Page 1

District Detail Report
District: [redacted]

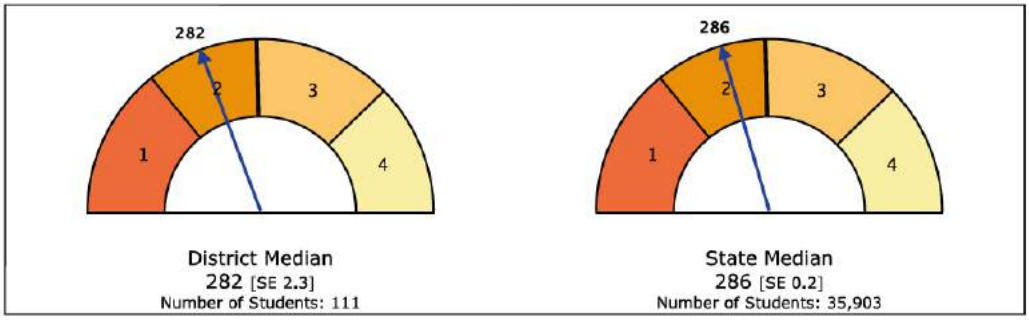
KANSAS ASSESSMENT PROGRAM

Grade 7 Mathematics
School Year: 2014–2015

This report has information about a district's scores from the Kansas Assessment Program. The tests measure students' understanding of Kansas College and Career Ready Standards at each grade using questions that ask students to select the right answer, sort items, create graphs, or label pictures. For sample test questions, see http://ksassessments.org/practice-tests.

District Median Score

The first graph shows the district's overall median score on the test, indicated by the arrow. The bands on the graph represent the four possible levels, with 4 being the highest level. The other graphs show the performance of seventh graders in the district and state. The median, or middle number in an ordered list of numbers, is used for these comparison graphs.

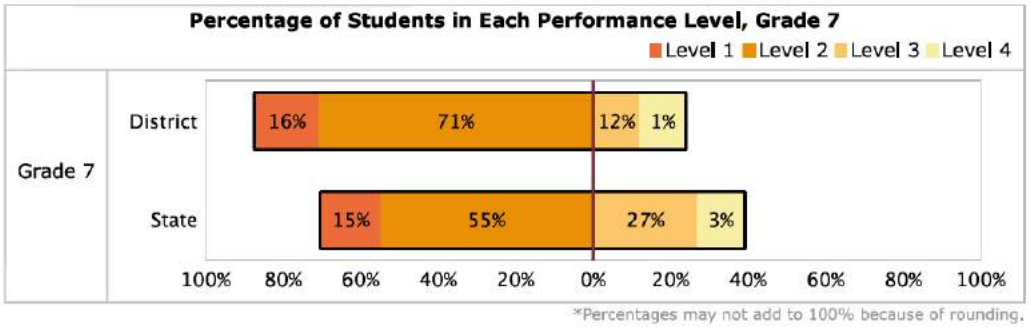


Performance Levels

Overall scores on the KAP test are divided into four performance levels. The levels range from 1 to 4, with 4 being the highest level. The district's median score is in Level 2.

Level	Score Range	Level Name
4	342 - 380	Level 4
3	300 - 341	Level 3
2	266 - 299	Level 2
1	220 - 265	Level 1

The following chart compares the percentage of seventh grade mathematics students in each performance level for the district and state. Complete performance level descriptors can be found at http://ksassessments.org/pld.



*Percentages may not add to 100% because of rounding.

Figure 5.10: DDR Page 2

District Detail Report [redacted] Grade 7 Mathematics

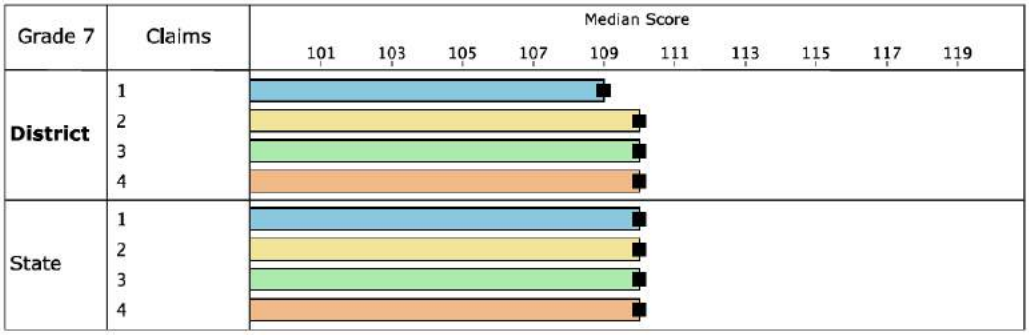
Explanation of Median and Standard Error

School, district, and state scores on this report are represented by the median score. A median is the middle number in an ordered list of numbers. For example, in the ordered list of scores {200, 210, 220, 230, 240, 250, 260}, the score of 230 is the median. The graphs show how the student's score compares to the median score for all students in the same grade who took the test in the school, district, and state.

Each score is also associated with a standard error of measurement (SE). The standard error around a student's score indicates how much a student's score might vary if the student took many equivalent versions of the test (a test with different items but covering the same content). The SE around the school, district, and state scores can be interpreted in a similar way. Standard error generally becomes smaller with larger comparison groups.

District Sub-Scores and Claims

This chart shows the district's performance on specific areas of the Grade 7 Mathematics test as well as the performance of the grade 7 students in the district and state. The bracket on either side of the bold score line represents the standard error, or how much a student's performance might vary if the student took many equivalent versions of the test.



Mathematics test questions cover four main areas (also called claims) of the Kansas Mathematics Standards.

- Claim 1: Concepts and Procedures. These questions require students to explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.
- Claim 2: Problem Solving. These questions require students to solve a range of complex problems using knowledge, problem solving strategies, and mathematical tools.
- Claim 3: Communicating and Reasoning. These questions require students to explain their reasoning, defend their answers, critique the reasoning of others and ask clarifying questions.
- Claim 4: Modeling and Data Analysis. These questions require students to analyze complex, real-world situations and construct and use mathematical models to solve problems, as well as interpret their result in the context of a situation.

Additional Resources

For the 2015 Interpretive Guide for score reports, visit http://kap.cete.us/ig.

5.7 Interpretative Guide

An *Interpretative Guide* for the reports is available on the KAP website.

See http://ksassessments.org/sites/default/files/documents/Interpretive%20Guide%20for%20Score%20Reports.pdf

5.8 Letter from the Commissioner of Education

An important part of the *Interpretive Guide* is a letter to Kansas educators and parents from Dr. Randy Watson, the Kansas Commissioner of Education. A copy of the letter is provided.

Figure 5.11: Commissioner's Letter to Educators and Parents

Thank you for supporting your student's participation in the Kansas Assessment Program. In Kansas, we believe in the need for high quality, meaningful assessments that are aligned to college and career ready academic standards and that challenge students to demonstrate the depths of their knowledge. The assessment your student took earlier this spring did just that.

While assessments should not be viewed as the "end all, be all," they do provide a critical piece of information that helps to inform instruction as well as provide consistent benchmarking to ensure students are prepared for whatever path they choose to pursue after graduation. State assessments provide an opportunity for teachers, parents, and students alike to check in on the student's progress.

Your student may have commented that this year's assessment was more difficult than in previous years, and they would be right. Kansas adopted more rigorous academic standards in 2010, and this year's assessment was the first time students were asked to demonstrate their mastery of skills such as critical thinking. When you receive your student's scores, it is important to remember they cannot be compared to your student's performance in previous years. Doing so would be like comparing apples to oranges – there simply is no comparison.

Kansas schools are among the best in the nation, and we all share in the responsibility of and commitment to ensuring the success of your student.

Thank you for your continued support of Kansas education and for being the most important champion for your student's education success.

Sincerely,
Dr. Randy Watson
Kansas Commissioner of Education

Kansas state department of Education

leadership and support for student learning

How can students improve their state assessment score?

- Talk with the classroom teacher about ways to develop your child's critical thinking skills.
- Ask your child questions that require explanations and can't be answered with a single word.
- Establish time for your child to read and provide suitable reading materials.
- Have your child write lists, letters, and other enjoyable or purposeful tasks.
- Solve math problems with your child using everyday materials such as road trip maps, sporting events, or recipes.
- Have your child explain to you how she or he solves math problems.

Because of the dramatic assessment format change as well as the increased rigor, results cannot be compared to previous assessments. The 2015 results will serve as a benchmark to measure future progress.

Part III

Assessment System Operations

6
Item Development

6.1 Item Development Process

ITEM WRITERS were trained in the use of KAP subject-area item specifications in the writing and reviewing of items. Items were reviewed by both AAI content experts and trained, external item reviewers. Before appearing on any assessment, items were reviewed by content reviewers, bias and sensitivity reviewers, and KSDE staff.

AAI staff used item-review feedback to revise test items as needed. Items and test forms were then prepared for field testing, according to test specifications. After field testing, item and test data were analyzed; this data analysis guided decisions about the use of items on future assessments.

6.2 Item Writing and Review

GRADUATE research assistants (GRAs) at University of Kansas were hired specifically to write items. They were hired based on their subject-matter expertise, prior item-writing experience, and teaching experience. GRAs who wrote items for the assessment majored in a variety of academic areas, including curriculum and teaching, mathematics, economics, pre-med, classical languages, biology, computer science, and earth and space sciences.

Before writing items for the KAP, all item writers received thorough in several topics, including:

- *Kansas College and Career Ready Standards* (KCCRS)
- validity and reliability
- alignment
- the difference between cognitive complexity and difficulty
- evidence-centered design (ECD)

- principles of universal design (UD) and accessibility
- bias and sensitivity
- item types (e.g., selected-response items, constructed-response items, technology-enhanced items)

Item-writing training also included extensive practice. Participants discussed depth of knowledge (DOK) for specific standards, examined practice items for alignment to those standards, and determined if practice items were written to the appropriate difficulty level. Participants also practiced writing items and received feedback from AAI staff.

Item writers received several documents to guide the item writing process. Item writers adhered to several guidelines that were made explicit in the training session. The guidelines are documented below.

6.2.1 Content Guidelines

- Write items to appropriate content standards.
- Ensure that multiple-choice items measure a single concept.
- Ensure that items focus on important ideas, not trivia.
- Use vocabulary that is consistent with students' grade level.
- Align items to the cognitive complexity of content standards.
- Write items to a variety of difficulty levels.

6.2.2 General Guidelines

- Write items that have clearly correct answer choice(s), with other answer choices wrong.
- Ensure that items are clearly worded.
- Avoid the use of tricky or misleading items.
- Proofread items for correct grammar, punctuation, and spelling.
- Avoid the use of contractions.
- Use third-person perspective.
- Avoid the use of humor.

6.2.3 Format Guidelines

- Format answer choices vertically rather than horizontally.
- Ensure that items use enough white space and are not cramped.
- Create clear layouts.
- Write clear instructions.

6.2.4 Structure Guidelines

- Avoid complex-format items.

- Write items in form of a question.
- Avoid *window-dressing* items (e.g., excessive verbiage).

6.2.5 Stem Construction Guidelines

- Write stems positively whenever possible.
- Avoid asking for and expressing opinions in stems.
- Ensure that the central idea is in the stem.
- Place the question as close to the answer choices as possible.
- Minimize the use of qualifying words (e.g., always, never).

6.2.6 Answer Choice Development Guidelines

- Order answer choices logically.
- Create independent answer choices that do not overlap.
- Write answer choices that are roughly the same length and parallel in the structure.
- Do not offer "all the above," "none of the above," or"I don't know" as answer choices.
- Avoid cluing between the stem and answer choices.
- Avoid specific determiners like "always" or "never."
- Create plausible distractors.
- Write strong distractors.
- Create distractors that take advantage of common errors and misconceptions.
- Answer keys should be roughly uniform in their distribution.

6.2.7 Accessibility Guidelines

- Consider access needs of special populations and how accommodations affect an item's intent.
- Use simple sentence structures.
- Minimize use of words with multiple meanings.
- Avoid the use of slang and regional dialect.
- Avoid the use of complicated names or names that could be confused with other nouns.
- Clearly label graphics.

6.2.8 Bias and Sensitivity Guidelines

- Avoid the use of stereotypes.
- Consider the regional and cultural nuances of words.
- Avoid the use of demeaning or offensive materials, particularly in the stimulus.

- Avoid the use of offensive or religious references.
- Ensure that items do not measure socioeconomic status or family attributes.
- Use artwork that reflects the diversity of the student population.

6.3 Item Reviews

THE ITEM REVIEW PROCESS involved several stages:

- Internal content review
- Psychometric review
- Accessibility review
- Editorial review
- KSDE review
- External content review, using multiple panelists
- External bias & sensitivity review, using multiple panelists
- Internal content team resolution, in consultation with KSDE

6.4 Item Reviewers

AAI content experts and KSDE staff selected two types of item reviewers: content reviewers and bias and sensitivity reviewers. Prospective item reviewers completed an online survey in which they indicated their demographic information, teaching experience, professional qualifications, content expertise, knowledge of the standards, and special education or ELL endorsements or training.

Content review panels were formed by grade band: grades 3–5, grades 6–8, and high school. Bias and sensitivity panels were assembled with members representing a number of minority groups. After completing a web-based training session, reviewers reviewed items at their own pace but before a given deadline. Item reviews were processed through a secure, online reviewing system.

6.4.1 Item Review Training

All reviewers completed two web-based item review training sessions: bias and sensitivity training and content review training. The training sessions included information about the KSDE–AAI partnership, test and item security, item writing guidelines, and the item review process. Item review training also provided participants with practice items and AAI staff contact information.

Bias and sensitivity reviewers were given a code sheet that provided coded categories and descriptions for possible concerns. For example,

possible gender bias was Code 1a and possible race or ethnicity bias was Code 1b. Descriptions and corresponding categories are given below.

- Possible bias related to gender, race or ethnicity, socioeconomic factors, or other
- Possible barrier related to uncommon or unfamiliar language, linguistic complexity or lack of clarity, assumed prior knowledge, cultural restrictions, accessibility, or other
- Possible sensitivity concern related to stereotype, religion, socioeconomic factors, status, specific topic, or other
- Other concern
- No barrier, bias, sensitivity, or other concern noted

For items flagged for evidence of bias or sensitive information, reviewers were instructed to use codes to provide details about why they flagged the item.

Content reviewers focused on the alignment of items to assessment targets, checking that items adequately addressed part of the target and elicited evidence for at least part of one evidence statement. Content reviewers also considered many other aspects of each item.

- Appropriate grade-level vocabulary
- Clear, complete statement or question
- Grammatically correct text
- Correct key
- Accurate, relevant graphics
- Well-designed answer choices that do not require background knowledge outside of the content area and free from *clang* associations

Clang occurs when words from an item's stem appear in one or more response options.

Based on their analysis of items, reviewers advised that items be accepted, revised, or rejected. Reviewers provided specific reasons to revise or reject items (e.g., "item does not align").

Prior to item writing, AAI also conducted reviews for ELA passages. Passage reviewers used qualitative and quantitative measures to examine text complexity and grade-level suitability . Reviewers also considered:

- Text length
- Sources of bias or sensitivity
- Possible overexposure (i.e., the text is commonly taught)
- Interest level
- Images
- Other elements considered significant

6.5 Universal Design in Test Development

DURING ITEM WRITER TRAINING, participants received instruction on UD concepts. Item writer training included a definition of UD and examples of test items that adhered to UD principles. Additionally, the general item writer guidelines (presented earlier in this chapter) included many UD principles.

6.6 Field Testing Process

THE 2015 ASSESSMENT was an *operational-pilot event*; a term used by AAI that suggests the hybrid nature of this testing event. Due to a cyber attack which shut down testing in 2014, psychometric data could not be gathered on the field-test (FT) items from the 2014 transitional assessment. Due to this, operational items on the 2015 tests could not be identified in advance of the 2015 testing window. As a mitigation strategy, the 2015 tests included an overage of test items, many of which were newly developed. When the 2015 testing window closed, item statistics were reviewed and the best performing items were selected as the operational set on which to base student scores.

Future assessments will include a block of 15 embedded FT items. Different sets of students will take different sets of FT items, resulting a larger pool of FT items, overall.

6.6.1 Field-Test Data Analysis

FT items are generally subjected to the same analyses as the operational items. The chapters on classical item analysis results and IRT calibration describe the statistics obtained for items using both of those modeling frameworks. DIF statistics are described in the chapter on fairness. Because there was no clear distinction between operational and FT items before the start of the spring 2015 testing event, only operational-item results are shared in this year's technical manual. In future events, the field-test items will be identified in advance of testing and will be placed in one of the blocks of items in the multistage tests that is specifically designated for FT items. Item statistics will be provided separately for operational and field-test items in future technical manuals.

7
Test Design and Development

The development of any test necessitates that many important decisions be made. These include deciding what content and cognitive levels (e.g., depth of knowledge) are important, as well as what scope, sequence, and progression of that content are appropriate for particular subject areas. Other decisions relate to the number of points needed for each test and the proportion of those points that are needed for any subscores to be reported. These decisions are not made in isolation but instead must be reasonable across all grade levels of the assessment. Together, all of these decisions help represent the construct(s) that a test measures.

7.1 Test Development Timeline

AAI worked with KSDE to determine the content to be assessed by the KAP subject-area and grade-level tests. The development leading up to the 2015 KAP test administration occurred over multiple years. The following table describes the test development timeline for both ELA and mathematics.

The KAP item development and review procedures are described in the prior chapter.

Table 7.1: Development Timeline

Milestone	Date	Notes
Adoption of Standards	October 2010	
Item Development	2011 to 2014 (ongoing)	
KCCR Item Types Included in Summative Assessment	Spring 2012 and spring 2013	Only machine-scorable items included to provide schools and districts a snapshot of performance on the KCCRS but not included in accountability measures
Transitional Summative Assessment	Spring 2014	Machine-scorable items only
Operational Pilot (Non-Adapting)	Spring 2015	Field testing performance tasks (not machine scorable)
Operational Stage-Adaptive Assessment	Spring 2016	Includes embedded field testing for machine-scorable items

7.2 Domain Sampling

Domain sampling[1] refers to the selection of a sample of test items from a well-defined population of items that define the domain of performance from which test score inferences will be made (e.g., college and career readiness in mathematics). At the heart of domain sampling is what Wainer[2] referred to as the *fundamental tenet of testing*:

[1] Crocker and Algina (1986)

[2] Wainer (2011)

> a relatively small sample of an individual's performance, measured under carefully controlled conditions, [can] yield an accurate picture of that individual's ability to perform under much broader conditions for a longer period of time.

In a compelling analogy, Koretz[3] compares test development to political polling. Koretz begins his analogy by pointing out that a poll for a national election might only sample 500 prospective voters to predict the voting outcomes from a population of 50 million people. That would mean that only one person is sampled by the poll for every 100,000 prospective voters in existence. As Koretz asserts, we don't care about the vote of any given individual in the poll's 500 person sample. Nor are the responses from all 500 persons in the poll important per se. The most important issue is the *generalization* that can be made from the 500 people to the 50 million prospective voters (i.e., do they prefer Candidate A or Candidate B). Because it is impossible to measure the thing that is really of interest—the preferred candidate among the 50 million voters—the proxy poll of 500 people is used instead.

[3] Koretz (2008)

In this analogy, a person sampled for the poll is analogous to an item sampled for a test.

It is common to hear that a political race is 'too close to call' or 'within a poll's margin of error.' In testing, error is an important consideration as well and is addressed in chapter on reliability

The very same thing is true in test development. We cannot administer all the items we want to, so we must rely on a small sample of items. Then, we generalize from performance on the sampled items to how students would have performed if they were given all items in the universe.

When all the key elements of a poll's development fit together, the poll's results can be very accurate. In testing, the adequacy of the assessment also depends on several important factors. Koretz (2008) notes the importance of motivated respondents in both polling and testing. Further, the wording of the questions has substantial effects on poll responses as well as test-taker responses. Of greatest emphasis for this chapter is the fact that the 500 people in the poll must represent the target population. Polling in a single geographic region, or a single gender, or a single ethnicity would yield misleading results. This is true in testing as well as the sampled items must adequately represent the domain of interest. Koretz calls this the *sampling principal of testing*, which is the same as Wainer's *fundamental tenant of testing*.

7.2.1 Sampling Philosophy for KAP

As noted above, test developers must address many issues when constructing a test. These include, among others, determining what content should be assessed and the number of items/points needed for each test score and each claim score reported. The KAP test developers understood that the addition of more items to any tests eventually has diminishing returns. Using their experience creating assessments, the number of total points allocated to the ELA and mathematics tests were determined so as to ensure those total test scores would be reliable for individual students.

Items were sampled at the claim level (which correspond to domains in the KCCRS Mathematics standards and ELA standards). Some claim scores will not be reliable at the student level, but may be reliable at the school level and at the right level in the KCCRS hierarchy where teachers can take meaningful instructional actions within their schools. This approach (1) keeps overall testing time to a minimum, and (2) provides students and educators with an appropriate breadth of information relative to the KAP's time demands.

7.3 Test Blueprints

THE BUILD TARGET IN MATH was 65%–75% in Claim 1, with the remaining items in Claims 2, 3, and 4 (roughly 8% to 12% in each). The ELA build target was 60%–65% in Claim 1 and the remaining items in Claim 2.

More specific content emphasis is included on the web. They are included in the appendix as well.

For ELA
http://www.ksassessments.org/sites/default/files/documents/ELA/Kansas%20Content%20Emphases%20for%20ELA.pdf.
For Math
http://www.ksassessments.org/sites/default/files/documents/Math/Kansas%20Content%20Emphases%20for%20Math.pdf.

7.4 Operational Test Construction

SOME RULES FOLLOWED during test construction including the following:

- Items placed in the first *block* (a 25 item set) were to be the ones with the best statistics based on field-test event in Kansas in the spring of 2014.
- Items in the subsequent blocks (three blocks each with 15 items) could have newly developed items. Although not previously administered, some of these items performed well enough that they contributed to student operational scores in 2015.
- The passages that had the most complete set of surviving items from the 2014 field-test event that matched the test blueprints were selected.
- For the items that had statistics from 2014, a wide range of item difficulties were chosen in order to achieve decent measurement precision across a wide range of student performance.
- Items and passages had to be acceptable to KSDE.
- Items on the test were reviewed to eliminate enemies (e.g., items that might give away answers to other items).
- In mathematics, the first set of 25 items were ordered from easiest to hardest based on their 2014 difficulty statistics.
- In ELA, the first set of 25 items were ordered according to the established protocol of KSDE and referencing order of appearance in the text.
- Items with the same key were to appear no more than three times in a row.

7.5 Characteristics of Final Test Forms

7.5.1 Content

KAP TESTS USE multiple test forms. In 2015 eight forms, each containing 70 items, were used for all ELA and math grade-level tests, corresponding to the future multistage adaptive testing (MST) format for KAP. This format will include a total of 70 items in four stages comprising 25, 15, 15, and 15 items in each stage.

Although there were 70 items, all of them were not used in calculating student test scores. After testing, item statistics were reviewed, and the items used in calculating student scores were selected on a form-by-form basis.

KAP wasn't able to designate operational items before the 2015 testing window opened, as is typically done, because of a cyberattack during the 2014 assessment year. This prevented KAP from determining the psychometric properties of many items. Consequently, operational items were selected after the close of the 2015 testing window. Item statistics were then available and were used to select the operational items. For these reasons AAI has referred to the 2015 event as an *operational pilot*.

Items were often excluded because of marginal performance statistics (e.g., poorly discriminating items), or the item statistics suggested that the items needed to be modified (e.g., the presence of multiple correct responses). Items were excluded for other reasons as well. To prepare for the transition to the MST administration format, some off-grade testing was done to expand item-pool difficulty. Such off-grade items were excluded this year, as were a few items that did not display properly on computer screens during testing.

Off-grade items might be used in the future but should measure an appropriate on-grade standard.

Because items were excluded on a form-by-form basis, some forms had more possible points than others. To avoid excluding additional items, the scores reported on each form were based on as many items as possible (in contrast to equalizing the possible points across forms to the minimum observed). The number of points and percentage of items allocated to claims varied slightly across forms. Specific results are documented in the following tables.

Table 7.2: Content Distribution for Grade 3 Math Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_2	Claim_3	Claim_4	Claim_234
Math	3	A	64	0.69	0.12	0.09	0.09	0.31
Math	3	B	51	0.71	0.10	0.12	0.08	0.29
Math	3	C	63	0.70	0.13	0.06	0.11	0.30
Math	3	D	66	0.68	0.11	0.11	0.11	0.32
Math	3	E	52	0.69	0.12	0.08	0.12	0.31
Math	3	F	61	0.72	0.10	0.08	0.10	0.28
Math	3	G	64	0.70	0.11	0.09	0.09	0.30
Math	3	H	63	0.70	0.13	0.08	0.10	0.30

k = Points Possible

Table 7.3: Content Distribution for Grade 4 Math Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_2	Claim_3	Claim_4	Claim_234
Math	4	A	67	0.67	0.10	0.10	0.12	0.33
Math	4	B	46	0.74	0.07	0.09	0.11	0.26
Math	4	C	67	0.69	0.10	0.10	0.10	0.31
Math	4	D	66	0.68	0.11	0.14	0.08	0.32
Math	4	E	51	0.71	0.12	0.08	0.10	0.29
Math	4	F	68	0.68	0.10	0.10	0.12	0.32
Math	4	G	67	0.69	0.10	0.13	0.07	0.31
Math	4	H	64	0.69	0.09	0.11	0.11	0.31

k = Points Possible

Table 7.4: Content Distribution for Grade 5 Math Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_2	Claim_3	Claim_4	Claim_234
Math	5	A	68	0.66	0.10	0.12	0.12	0.34
Math	5	B	52	0.69	0.10	0.10	0.12	0.31
Math	5	C	65	0.68	0.09	0.12	0.11	0.32
Math	5	D	65	0.69	0.09	0.11	0.11	0.31
Math	5	E	51	0.73	0.08	0.10	0.10	0.27
Math	5	F	62	0.71	0.08	0.10	0.11	0.29
Math	5	G	64	0.69	0.11	0.11	0.09	0.31
Math	5	H	65	0.68	0.12	0.11	0.09	0.32

k = Points Possible

Table 7.5: Content Distribution for Grade 6 Math Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_2	Claim_3	Claim_4	Claim_234
Math	3	A	64	0.69	0.12	0.09	0.09	0.31
Math	3	B	51	0.71	0.10	0.12	0.08	0.29
Math	3	C	63	0.70	0.13	0.06	0.11	0.30
Math	3	D	66	0.68	0.11	0.11	0.11	0.32
Math	3	E	52	0.69	0.12	0.08	0.12	0.31
Math	3	F	61	0.72	0.10	0.08	0.10	0.28
Math	3	G	64	0.70	0.11	0.09	0.09	0.30
Math	3	H	63	0.70	0.13	0.08	0.10	0.30

k = Points Possible

Table 7.6: Content Distribution for Grade 7 Math Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_2	Claim_3	Claim_4	Claim_234
Math	7	A	60	0.70	0.12	0.12	0.07	0.30
Math	7	B	47	0.74	0.11	0.11	0.04	0.26
Math	7	C	63	0.70	0.13	0.11	0.06	0.30
Math	7	D	62	0.71	0.13	0.10	0.06	0.29
Math	7	E	51	0.69	0.12	0.10	0.10	0.31
Math	7	F	65	0.69	0.12	0.09	0.09	0.31
Math	7	G	61	0.69	0.11	0.10	0.10	0.31
Math	7	H	48	0.67	0.12	0.10	0.10	0.33

k = Points Possible

Table 7.7: Content Distribution for Grade 8 Math Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_2	Claim_3	Claim_4	Claim_234
Math	8	A	63	0.73	0.10	0.08	0.10	0.27
Math	8	B	49	0.71	0.10	0.08	0.10	0.29
Math	8	C	56	0.71	0.07	0.11	0.11	0.29
Math	8	D	61	0.75	0.07	0.08	0.10	0.25
Math	8	E	48	0.75	0.08	0.10	0.06	0.25
Math	8	F	59	0.75	0.10	0.08	0.07	0.25
Math	8	G	61	0.72	0.07	0.11	0.10	0.28
Math	8	H	47	0.74	0.06	0.11	0.09	0.26

k = Points Possible

Table 7.8: Content Distribution for Grade 10 Math Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_2	Claim_3	Claim_4	Claim_234
Math	10	A	57	0.74	0.09	0.11	0.07	0.26
Math	10	B	55	0.67	0.15	0.09	0.09	0.33
Math	10	C	59	0.69	0.14	0.07	0.10	0.31
Math	10	D	58	0.74	0.10	0.10	0.05	0.26
Math	10	E	57	0.74	0.09	0.07	0.11	0.26
Math	10	F	56	0.71	0.11	0.09	0.09	0.29
Math	10	G	60	0.75	0.08	0.08	0.08	0.25
Math	10	H	52	0.67	0.13	0.08	0.12	0.33

k = Points Possible

Table 7.9: Content Distribution for Grade 3 ELA Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_1_Info	Claim_1_Lit	Claim_2
ELA	3	A	74	0.61	0.28	0.32	0.39
ELA	3	B	58	0.67	0.36	0.31	0.33
ELA	3	C	73	0.62	0.27	0.34	0.38
ELA	3	D	58	0.66	0.36	0.29	0.34
ELA	3	E	77	0.65	0.29	0.36	0.35
ELA	3	F	76	0.61	0.26	0.34	0.39
ELA	3	G	76	0.64	0.41	0.24	0.36
ELA	3	H	74	0.65	0.28	0.36	0.35

k = Points Possible

Table 7.10: Content Distribution for Grade 4 ELA Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_1_Info	Claim_1_Lit	Claim_2
ELA	4	A	80	0.64	0.25	0.39	0.36
ELA	4	B	58	0.71	0.36	0.34	0.29
ELA	4	C	71	0.69	0.41	0.28	0.31
ELA	4	D	60	0.67	0.33	0.33	0.33
ELA	4	E	74	0.62	0.36	0.26	0.38
ELA	4	F	77	0.64	0.26	0.38	0.36
ELA	4	G	72	0.67	0.42	0.25	0.33
ELA	4	H	69	0.68	0.29	0.39	0.32

k = Points Possible

Table 7.11: Content Distribution for Grade 5 ELA Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_1_Info	Claim_1_Lit	Claim_2
ELA	5	A	76	0.63	0.26	0.37	0.37
ELA	5	B	61	0.66	0.34	0.31	0.34
ELA	5	C	74	0.64	0.43	0.20	0.36
ELA	5	D	61	0.64	0.33	0.31	0.36
ELA	5	E	77	0.62	0.38	0.25	0.38
ELA	5	F	71	0.65	0.41	0.24	0.35
ELA	5	G	77	0.60	0.26	0.34	0.40
ELA	5	H	75	0.64	0.39	0.25	0.36

k = Points Possible

Table 7.12: Content Distribution for Grade 6 ELA Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_1_Info	Claim_1_Lit	Claim_2
ELA	6	A	73	0.64	0.27	0.37	0.36
ELA	6	B	57	0.63	0.33	0.30	0.37
ELA	6	C	69	0.65	0.42	0.23	0.35
ELA	6	D	58	0.64	0.36	0.28	0.36
ELA	6	E	73	0.59	0.25	0.34	0.41
ELA	6	F	72	0.64	0.29	0.35	0.36
ELA	6	G	69	0.68	0.28	0.41	0.32
ELA	6	H	66	0.64	0.39	0.24	0.36

k = Points Possible

Table 7.13: Content Distribution for Grade 7 ELA Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_1_Info	Claim_1_Lit	Claim_2
ELA	7	A	72	0.68	0.28	0.40	0.32
ELA	7	B	58	0.67	0.38	0.29	0.33
ELA	7	C	74	0.66	0.39	0.27	0.34
ELA	7	D	58	0.60	0.36	0.24	0.40
ELA	7	E	80	0.61	0.24	0.38	0.39
ELA	7	F	82	0.59	0.35	0.23	0.41
ELA	7	G	80	0.59	0.22	0.36	0.41
ELA	7	H	76	0.59	0.38	0.21	0.41

k = Points Possible

Table 7.14: Content Distribution for Grade 8 ELA Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_1_Info	Claim_1_Lit	Claim_2
ELA	8	A	75	0.63	0.25	0.37	0.37
ELA	8	B	57	0.65	0.32	0.33	0.35
ELA	8	C	68	0.63	0.28	0.35	0.37
ELA	8	D	55	0.69	0.38	0.31	0.31
ELA	8	E	64	0.61	0.42	0.19	0.39
ELA	8	F	76	0.63	0.28	0.36	0.37
ELA	8	G	73	0.60	0.23	0.37	0.40
ELA	8	H	70	0.59	0.26	0.33	0.41

k = Points Possible

Table 7.15: Content Distribution for Grade 10 ELA Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_1_Info	Claim_1_Lit	Claim_2
ELA	10	A	71	0.62	0.27	0.35	0.38
ELA	10	B	74	0.69	0.39	0.30	0.31
ELA	10	C	72	0.71	0.29	0.42	0.29
ELA	10	D	85	0.60	0.24	0.36	0.40
ELA	10	E	71	0.59	0.25	0.34	0.41
ELA	10	F	82	0.56	0.33	0.23	0.44
ELA	10	G	83	0.58	0.33	0.25	0.42
ELA	10	H	83	0.60	0.25	0.35	0.40

k = Points Possible

7.5.2 Psychometric Properties

The second half of this technical manual reviews the psychometric properties of the final test forms. Of particular interest is the chapter on operational test statistics.

Because of cyberattacks during the 2014 assessment year, the first administration of the KAP was an *operational–pilot* event. As such, there were no specific, statistical build targets for 2015. This year's assessments will guide the construction of next year's tests; the statistical targets used during test construction, such as summary item

statistics and test characteristic curves will be included in future technical manuals.

7.6 Alignment

An independent alignment study was conducted by edCount, LLC®. Results of this study are available in a seperate report.

8
Test Administration

8.1 Test Sessions, Sections, Ticketing, and Timing

The KAP Math and ELA tests were administered in separate test sessions. The math tests had five sections in Grades 3–8 and four sections in Grade 10. The ELA tests had four sections at every grade levels. Each section required a separate *ticket*, and the first test section had to be completed before any subsequent tickets were made available.

One less section was required in Grade 10 math because there was no performance task at that grade.

The testing *ticket* confirms that a student is registered to take a particular test.

Each student was allowed as much time as necessary to complete each test session in one sitting. However, educators were provided time estimates for each section of both the Math and ELA tests: 50 minutes for Section 1; 25 minutes for Section 2; and 25 minutes each for Section 3 and 4. The expected time for the math performance task was 50 minutes.

8.2 Test Layout

Soft materials, such as test directions, were presented to students at the start of the math and ELA tests. Soft breaks in the math tests generally occurred between calculator-active and calculator-inactive portions of the test. Soft breaks in ELA generally occurred between reading passages.

8.3 Testing Window

The spring 2015 testing windows were February 16 - March 10 for the mathematics performance tasks and March 16 - April 28 for the machine-scorable portions of the ELA and math tests.

8.4 Testing Platforms

KITE is the testing platform for the assessments. The KITE website includes detailed information about the KITE system.

http://www.ksassessments.org/kite

8.5 Online System Usage During Testing Window

Google Analytics[1] was used to monitor several performance metrics of the KITE system during testing. The following figures and tables provide examples of the metrics that are tracked.

[1] ©2015 Google Inc.

Figure 8.1: System Usage by Date

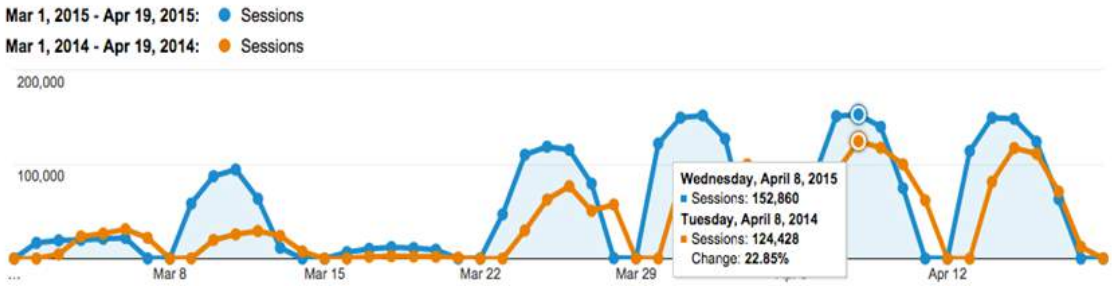


Table 8.1: Metrics for Two Test Administration Years

Metric	Spring 2014	Spring 2015	% Change
Session	1,705,358	2,730,360	60%
Users	444,323	837,567	89%
Avg Session Duration	09:00.0	37:00.0	22311%
Pageviews	5,165,114	110,771,350	2045%

Figure 8.2: Browser usage Pie Chart

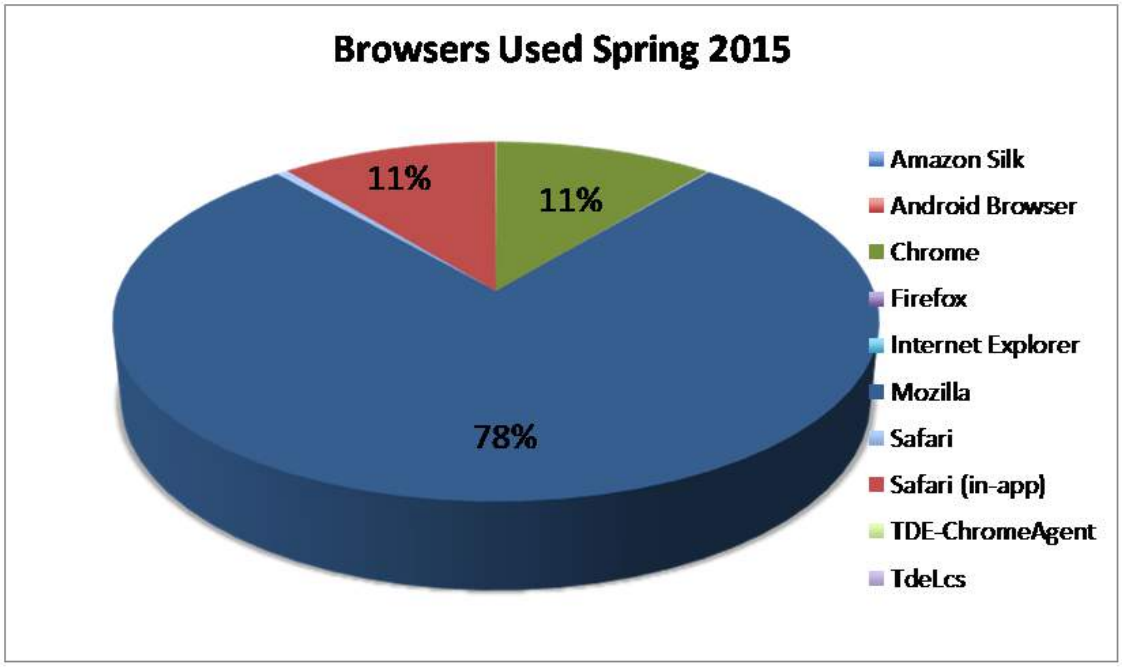


Table 8.2: Browser usage Table

Browser	Spring 2014	Spring 2015	% Change
Amazon Silk	4	2	-50%
Android Browser	33	18	-45%
Chrome	15,641	297,782	1804%
Firefox	16,928	2,509	-85%
Internet Explorer	8,464	518	-94%
Mozilla	1,459,548	2,119,652	45%
Safari	136,502	15,629	-89%
Safari (in-app)	68,183	293,201	330%
TDE-ChromeAgent	-	883	100%
TdeLcs	-	162	100%

8.6 Availability of Score Reports

SCORE REPORTS were made available through the KITE system in the fall of 2015. Due to standard setting and need for the state board to approve cut scores for performance levels, the reports were delayed in 2015. Moving forward, score reports will be available in the summer.

See the chapter on score reporting for additional information.

8.7 Technical Support

KITE HELP DESK was available via email (kap_support@ku.edu) or phone (855-277-9752) and was staffed from 8:00 a.m. to 8:00 p.m.

8.8 Ongoing Quality Control (QC) in Test Administration

SHORT-TERM QC focuses on quickly solving issues brought to the attention of the help desk. When required, fixes for one testing site are applied universally to prevent similar problems from occurring elsewhere. Issues are usually resolved within 24 hours.

Long-term QC focuses on resolving issues aired during annual lessons-learned meetings after each testing window.

9
Test and Data Security

Detailed information about test security is also outlined in the *Kansas Assessment Examiner's Manual*.

9.1 Test Administrator Training on Security

All district coordinators receive annual training on test security procedures. Training was provided during the pre-conference at the October 26th KSDE Annual Conference. Additionally, online training materials are provided on the KSDE assessment website. District test coordinators trained building-level personnel before local testing[1].

[1] *Kansas Assessment Examiner's Manual*, 2015 - 2016, p. 3

9.2 Test Security Plan

All district coordinators must be trained annually on test security procedures. Training will be provided during the pre-conference at the October 26th KSDE Annual Conference and via online training materials on the KSDE assessment website. District test coordinators will train building-level personnel before local testing.

All local personnel who administer state assessments must read the *Kansas Appropriate Testing Practices Fact Sheet*, found on the KSDE website. Local personnel must sign an agreement to abide by state ethical testing practices. The *Agreement to Abide by Guidelines* on Page 34 of the *Kansas Assessment Examiner's Manual* may be used for this purpose.

Building coordinators should submit testing schedules to district coordinators. KSDE staff will contact schools and districts by phone or email to schedule monitoring visits.

Using a form similar to the sample used in test security training sessions, district test coordinators must maintain documentation of all

accommodation needs. District test coordinators do not need copies of IEPs, 504 plans, or SIT plans.

KSDE staff and members of the Kansas Assessment Advisory Council will visit and evaluate five percent (5%) of Kansas schools during test administration. The evaluation checklist will be available on the KSDE website.

9.3 Guidelines

Test security guidelines for 2015 follow.

9.3.1 Guidelines for Educators

- Teachers must read the *Kansas Appropriate Testing Practices Fact Sheet* found on the KSDE website.
- Teachers must be trained in test security procedures.
- KSDE employees and members of the Kansas Assessment Advisory Council will visit five percent (5%) of schools administering state assessments. KSDE will post the checklist used by KSDE during test administration visits on the KSDE website.
- Teachers must immediately report any breach of test security or loss of materials to the building or district test coordinator.
- Teachers are responsible for the security of test materials. Paper copies of test booklets and tickets should be kept in a secure, locked area before, between, and after each testing session.
- The teacher is responsible for collecting and destroying (e.g., shredding, burning) student notes, scratch paper, and drawings at the end of each testing session.
- Teachers may not review tests or analyze test items before, during, or after the assessment is administered.
- Teachers and administrators may not keep copies of the tests.
- Teachers may not discuss with students any specific test items before, during, or after testing.
- Teachers may not copy, reproduce, or paraphrase test materials, nor may they construct parallel questions or cloned questions from actual test items.
- Teachers and administrators must strictly follow KSDE procedures concerning eligible students and the procedures required for administering read-alouds. These procedures will be monitored via KSDE visits.

9.3.2 Guidelines for Students

- Students may use blank paper to show and check their work. The paper must be collected and destroyed at the end of the test session.
- After finishing the test, students may not return to previously completed test sections unless items have been skipped or omitted.
- Students may create graphic organizers on blank paper during the test.
- Students may not use electronic devices (including cell phones, tablets, and similar devices) on any portion of the assessment.
- Students may use scratch paper, graph paper, and manipulatives on the mathematics assessments. However they may not use textbooks, dictionaries (with the exception of bilingual translation dictionaries for ELL students), and other curricular materials during testing.
- Students should make up any test session that occurred when the student was absent.

9.4 Ethical Issues for Educators

- During testing, teachers may not respond to questions that would help students understand a question, help them respond to an item, or advise or encourage students to edit or change a response.
- During testing, teachers may not direct or prompt students to use specific strategies. They may, however, remind students of strategies the day before testing.
- Teachers must complete review sessions on content and test-taking strategies before testing.
- Teachers may not coach or cue students in any way during testing, including the use of gestures or facial expressions.
- Teachers may not say nor do anything that would clue students about the accuracy of an answer or provide any advantage during testing.
- Teachers may not ask students how they reached an answer.
- Teachers may not tell, prompt, or hint that students should review a question or portion of the test.
- Teachers may not provide students the meanings of words in the text or in questions.
- Teachers may not read the reading passages on the ELA test. However, teachers may read isolated words, phrases, or sentences from any parts of any other assessments.
- Teachers may not review, teach, or practice tested indicators after

These are, at a minimum, ethical issues. Some may be better viewed as direct threats to the integrity of KAP test data.

content-area testing has begun.

- Teachers may not construct answer keys, and assessments may not be scored locally.
- Teachers may not require students to use scratch paper, show their work, or use the online tools (e.g., the highlighter tool).
- Teachers may not grade scratch paper.
- Teachers may reactivate students who have omitted items or who did not finish a test session due to illness, time constraints, or other factors. Teachers should monitor these students to ensure they work only on previously omitted items.
- Teachers may not reactivate, due to poor test performance or lack of effort, students who have completed all test sessions. If a student appears not to be trying or to be quickly clicking through answer choices, teachers should stop the test and contact the building test coordinator.

9.5 FERPA

For convenience, information about FERPA is repeated here although it also appears elsewhere in this technical manual.

The Family Educational Rights and Privacy Act (FERPA) affords parents and all students certain privacy rights with respect to their educational records. These rights extend to students in all grade levels, beginning in preschool and extending through their years in postsecondary institutions. The law applies to all schools that receive funds under any applicable program of the U.S. Department of Education.

Staff who work on the KAP that must have access to student data for their work take the protection of student privacy and confidentiality seriously. KAP data records are maintained on a secure server. Access to KAP data is limited according to the specific job responsibilities of staff. Staff with access to that data undergo yearly training on FERPA. Certification of all staff member's completion of FERPA training is maintained by the FERPA compliance officers.

9.6 Possible Security Enhancements for the Future

Kansas's technical advisory committee (TAC) advises the KAP on security and other technical aspects of the program. In the future, a likely test-design enhancement that will improve security will involve the use of multiple, multistage test panels once the item pool is deep enough to support that change. This step minimizes the exposure of items to testing students. Other enhancements will be employed after further discussions with TAC.

10
Materials Handling and Processing

10.1 Shipping, Packaging, and Delivery of Materials (Paper and Braille Tests only)

Questar, a subcontractor for AAI, delivered Braille booklets to the KAP Test Operations team at AAI. The team then packaged and delivered the booklets to schools according to the student personal-needs profile (PNP) information contained in KITE's Educator Portal. As noted earlier, fewer than 50 students per grade took the KAP outside the KITE testing engine.

10.2 Materials Storage and Return

The building test coordinator, when designated, was responsible for maintaining test-security documentation. Building test coordinators also maintained the security of Braille forms by delivering the forms to test administrators within 24 hours of test administration. Tests were not to be copied or taken out of the building, and the building test coordinator also returned all tests to the district test coordinator immediately after completion of the assessment.[1]

[1] *Kansas Assessment Examiner's Manual*, 2015 - 2016, p. 21

10.3 Manual

The *Kansas Assessment Examiner's Manual* (KAEM) can be found on the KAP website. Only the most up-to-date version of the KAEM is maintained on the website.

http://www.ksassessments.org/news

10.4 Test Administrator Training

DISTRICT COORDINATORS receive annual training on materials handling and security procedures. Training was provided during the preconference at the October 26th KSDE Annual Conference. Additionally, online training materials are provided on the KSDE assessment website. District test coordinators trained building-level personnel before local testing[2].

[2] *Kansas Assessment Examiner's Manual*, 2015 - 2016, p. 3

Part IV

Inclusion of All Students

11
Demographics, Inclusion, and Participation

This chapter presents information about KAP's inclusion policies. In addition, summary statistics are provided for key demographic groups as well as participation rates.

The demographic summary statistics will be monitored on an ongoing basis to detect unusual changes that might warrant further investigation.

11.1 KAP Inclusion Policy

Kansas is committed to including all students in the KAP assessments. Some notable exceptions that occur in Kansas include:

- If a student was serving a long-term suspension
- If a student was truant more than two consecutive weeks at time of testing
- If a student had a catastrophic illness or accident
- If a student moved during testing
- If a student was incarcerated

These exclusions are fairly typical across many state testing programs.

11.1.1 Procedures for Including English Language Learners

In addition to the exceptions above, English language learners (ELLs) who were recent arrivals to the United States took the KAP mathematics tests, but they only count toward participation. Further, ELL students were to take the KELPA-P in lieu of KAP ELA.

11.1.2 Procedures for Including SWDs

One percent of Kansas students took the DLM alternate assessment. Other special needs students with an IEP, 504, or SIT plans took the KAP tests but had accommodations consistent with their personal-needs profile (PNP). If an accommodation was given that

additional information is provided in the accmmodations chapter.

was not allowed (e.g., reading to student on the KAP ELA test) the student was treated as *not tested*.

11.2 Participation Data

The following participation data was provided by the KSDE. Seperate results are provided by subject area but aggregated over all grade levels. ELA and math results were very similar. The Valid Participation n-count is very similar total n count in all cases.

Table 11.1: Participation by Student Group for Math

Subgroup	Total	Valid Part.	Not Tested	Part.
All Students	244,465	243,351	1,360	241,991
Free and Reduced Lunch	118,827	118,197	764	117,433
Students with Disabilities	32,209	31,958	395	31,563
ELL Students	28,654	28,575	13	28,562
African-American Students	17,462	17,237	267	16,970
Hispanic	45,904	45,698	153	45,545
White	159,654	159,101	841	158,260
Asian	6,616	6,593	9	6,584
American Indian or Alaska Native	2,317	2,298	12	2,286
Multi-Racial	12,119	12,031	77	11,954
Native Hawaiian or Pacific Islander	393	393	1	392

Table 11.2: Participation by Student Group for ELA

Subgroup	Total	Valid Part.	Not Tested	Part.
All Students	244,378	243,171	1,360	241,811
Free and Reduced Lunch	118,782	118,084	781	117,303
Students with Disabilities	32,215	31,957	390	31,567
ELL Students	28,544	28,361	19	28,342
African-American Students	17,460	17,218	275	16,943
Hispanic	45,863	45,587	158	45,429
White	159,640	159,089	832	158,257
Asian	6,587	6,559	12	6,547
American Indian or Alaska Native	2,316	2,299	11	2,288
Multi-Racial	12,119	12,028	71	11,957
Native Hawaiian or Pacific Islander	393	391	1	390

11.3 Participation by Administration Mode

Participation/nonparticipation by test administration mode is not available. However, the chapter on accommodations does provide n-counts by testing mode. By far the most common testing mode was computer. Very few students took the KAP in a different test administraiton mode.

11.4 Student Demographics

The demographic profile of students for math and ELA are presented separately but are extremely similar, as expected. There was some fluctuation in the proportion of students across grades, but the difference was generally no more than 0.02 to 0.03 in value.

The proportion of White students in Grade 6 differed between math and ELA.

During this test administration, the largest racial group was White, which constituted about four-fifths of students. African American students made up about 10% of the population. Regarding ethnicity, about one-fifth of the students were Hispanic. ESOL (English speakers of other languages) students made up just over 10% of the population as did SWDs (students with disabilities).

Table 11.3: Proportion of Students in Demographic Groups by Grade for Math

Group by Grade	3	4	5	6	7	8	10
AfricanAmerican	0.08	0.07	0.07	0.07	0.07	0.07	0.07
AmericanIndian	0.03	0.04	0.04	0.04	0.04	0.04	0.04
Asian	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Multi	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NativeHIorPacIsl	0.00	0.00	0.00	0.00	0.00	0.00	0.00
White	0.80	0.80	0.79	0.79	0.79	0.80	0.81
Hispanic	0.20	0.19	0.19	0.19	0.18	0.18	0.17
ESOL	0.13	0.13	0.13	0.12	0.12	0.11	0.09
SWD	0.12	0.12	0.12	0.12	0.12	0.11	0.10

Group by Grade	3	4	5	6	7	8	10
AfricanAmerican	0.08	0.07	0.07	0.07	0.07	0.07	0.07
AmericanIndian	0.03	0.04	0.04	0.04	0.04	0.04	0.04
Asian	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Multi	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NativeHIorPacIsl	0.00	0.00	0.00	0.00	0.00	0.00	0.00
White	0.80	0.80	0.79	0.80	0.79	0.80	0.81
Hispanic	0.19	0.19	0.19	0.19	0.18	0.18	0.17
ESOL	0.13	0.13	0.13	0.12	0.11	0.11	0.09
SWD	0.12	0.12	0.12	0.12	0.12	0.11	0.10

Table 11.4: Proportion of Students in Demographic Groups by Grade for ELA

12
Accommodations

This chapter presents information about KAP accommodation usage. The types of accommodations available for KAP are reviewed as well as rules for their use. Although much of this information is available in other KSDE documents (i.e., *Tools and Accommodations for the Kansas Assessment Program (KAP) 2014 – 2015* and the *Kansas Assessment Examiner's Manual 2014 – 2015*) it is recaptiulated below for the reader's convienence. The chapter closes with a summary statistics are provided regarding frequency of use of the accommodations as well.

12.1 General Overview

There are a few basic rules that are common for every available tesitng accommodation on KAP. First and foremost, accommodations should not be used on the state assessments if they have not been a regular part of instruction. Second, IEP, 504 and SIT students may only use accommodations that are documented in their IEP, 504, and SIT plans. Finally, to be used during the KAP, the accommodations must have been documented in the student's personal-needs profile (PNP) in Educator Portal.

12.2 Prohibited Accommodations

Some accommodations are prohibited for all students. In general, reading to students any text (including isolated words) in the passages on the English language arts (ELA) test is prohibited. Violations will result in the student being counted as *not tested*. For a very limited number of students, such as those who could not access printed text due to blindness or low vision and did not have adequate Braille skills, text-to-speech (TTS) reading of reading passages was permitted.

Students were not permitted to use their own calculators with the

exception of accommodated mathematical tools for students with disabilities. For example, a students with a documented need for special mathematical tools (such as an abacus or large button calculator for visually impaired students) were permitted to use the tool as documented in their IEP. Students could also use handheld calculators (vs. KITE's calculator tool) on calculator items as an accommodation if it was documented in the student's IEP, 504, or SIT plan.

Other prohibited accommodations included:

Use of teacher-generated or student-generated journals, notes, and logs is prohibited. * Use of commercially-made, teacher-made, or teacher-generated graphic organizers is prohibited.

12.3 Accommodations for ELL Students

For convenience, KAP may be given in small groups of no more than three ELL students. Test directions may be read to the student in English or explained in the student's native language. Electronic translators and bilingual dictionaries may be used for directions, test questions and answer choices. However, they cannot be used on any ELA passages. Further, English language arts passages, test questions, answer choices, labels, graph titles, or other items may NOT be translated into the student's native language.

Spanish versions of the mathematics assessment is available. Only students who have been instructed in Spanish may use Spanish versions of the tests. The entire mathematics assessment may be read to students in the Spanish version. However, the adult reader is *not* allowed to translate from one language to another. It was possible for some students to have both an individual read-aloud accommodation and the paper/pencil accommodation in Spanish (available in mathematics only) when the student needed to respond on a paper form while an adult reads from a Spanish paper test.

12.4 Paper/Pencil Accommodation

All students were to take the KAP tests on computer except in very rare circumstances in which a paper/pencil accommodation were allowed. This accommodations allowed PDF test documents to be the printed through KITE'S Educator Portal as an individual accommodation. As with other accommodations, this accommodation must routinely be used in the classroom (e.g., when other students are using the computer the student in need of the accommodation users paper). There are several key questions to ask about the child when considering a paper/pencil accommodation:

paper/pencil tests could not be requested for entire classes.

- Has the student used the computer for formative assessment(s)?

- Does the student have barriers to using the computer in individual or group instructional settings that require alternative assignments when the class is using the computer?

As with other accommodations, a student's need for the paper/pencil accommodation must be documented in a pre-intervention plan (student improvement plan), ELL plan, 504 plan, or IEP. The following must be documented must be included in those plans:

- Student name
- Student grade
- Building/district name
- Evidence documenting the need for the paper/pencil accommodation.

Evidence suggesting need for the accommodation could possibly include (1) progress monitoring data, (2) English language arts level of instructional materials used in classroom, (3) documentation that the paper/pencil accommodation is used in the classroom setting for both instructional materials and assessments. The implementation date of the classroom accommodation must be listed, (4) signatures of team members involved in the decision to recommend the paper/pencil accommodation including the student's teacher and building administrator, (5) no answer sheets may be generated by the school or district. Students must mark their answers on the paper copy of the assessment, and (6) district or building-level personnel will work in pairs to enter student answer choices into KITE. Documentation of the need for paper/pencil accommodations must be kept on file by the district test coordinator. KSDE staff will monitor 5% of all test administration sessions and will ask at each monitoring visit to see documentation of paper/pencil and read-aloud accommodations.

12.5 Recording Accommodations

Students receiving read-aloud via headphones could take the test in the typical group setting. Schools and districts could continue to use the following accommodations, without a need to report them to KSDE:

- Separate, quiet, or individual setting
- Frequent breaks
- Student dictation of answers to scribe
- Student use of communication device
- Some other accommodation was used
- American Sign Language delivery of directions to student

- Student response in American Sign Language
- Student use of Braille writer or slate and stylus
- Student self-reading aloud of assessment
- Student use of translation dictionary

12.5.1 Text-to-Speech (TTS) Accommodations Policy

The TTS accommodation is for a student who needs the entire assessment read aloud (with the aforementioned exception of ELA passages). A student who needs a TTS accommodation is one whose ability to convey knowledge of the subject/content area is severely limited by his/her inability to read the assessment materials. The student cannot or would not be successful in the classroom without the read-aloud accommodation. To use the TTS accommodation on the state assessment, the student must have the read-aloud accommodation provided in the classroom on a regular basis (i.e., as an on-going practice for both classroom instruction and classroom assessments/tests). Neither English language learners nor students who receive Title I or special education services automatically qualify for the TTS accommodation.

From the above, it is the local district's responsibility to define what is meant by *severely limited* and to quantify what a *regularbasis* is for classroom instruction and assessments/tests. Tools for determining need and resources available may be selected by individual districts. However, the general expectation is that students will be more than one year below grade in ELA and that the accommodation is being systematically applied at least 50% of the time on classroom assignments and 100% of the time on classroom assessments contributing to classroom grades.

12.5.1.1 Documenting the Need for TTS

Documentation needed for TTS is the same as paper/pencil accommodation reviewed above.

12.5.2 Allowable Practices

The read-aloud accommodation does not refer to an adult reading an occasional word, an occasional distractor, an occasional stem, or an occasional question to the student. In fact, practices such as pronouncing an occasional word, an occasional distractor, an occasional stem, or an occasional question should be considered acceptable assessment practice that doesn't require special documentation. Educators should use professional discretion regarding the number of times a student may request assistance. Teachers, test administrators, and

proctors may not read anything aloud for a student from an ELA passage. Such a practice is prohibited save for one exception. As noted earlier, a very limited number of non-visual students may have the passages read aloud using the KITE TTS feature.

12.5.3 KITE Text-to-Speech (TTS) Features

A synthetic voice is provided in KITE. The student's PNP in Educator Portal must indicate Spoken Audio with Synthetic marked under Voice Source. When *Read At Start* is set to False, the student must click or tap the play button for the read-aloud to begin. The *Spoken preference* option indicates which elements of a question should be read to the student. When reading items, the KITE audio voice will read the entire text of the question stem and answer options to students. To adjust the volume, the student will need device controls. They should set the audio volume prior to launching KITE Client. *Audio for directions only* should be set to False. Selecting True will only permit the directions to be read aloud to students, not the items. For students who receive daily instruction orally and through computerized text-to-speech systems and who have received permission from KSDE for the reading passages to be read aloud, select NonVisual under *Spoken Preference*.

12.6 Frequency of Accommodations Use

The following tables show the frequency of accommodations use on the KAP.

Table 12.1: N-counts for Various Personal Need Profiles by Grade for Math

PNP	3	4	5	6	7	8	10
Auditory Background	59	48	33	21	18	11	11
Background Colour	8	21	10	6	6	13	2
Braille	3	5	2	6	2	2	1
Colour Overlay	28	45	43	16	11	7	1
Foreground Colour	8	21	10	6	6	13	2
General	32,433	30,683	24,093	29,519	32,672	18,247	33,480
Invert Colour Choice	6	8	9	2	10	4	0
Item-Translation Display	11	18	20	27	35	28	50
Keyword Translation Display	6	10	8	7	14	8	48
Large-Print Booklet	10	16	7	7	8	7	2
Masking	16	54	22	30	55	5	3
Onscreen Keyboard	28	19	14	7	11	14	0
Paper and Pencil	24	20	10	16	25	16	3
Signing	13	11	1	1	9	11	14
Spoken	3,696	3,976	3,162	3,072	3,237	1,672	1,463
Tactile	1	3	0	0	0	2	0
Total N	36,213	34,719	27,304	32,632	36,003	19,961	35,008

Table 12.2: N-counts for Various Personal Need Profiles by Grade for ELA

PNP	3	4	5	6	7	8	10
Auditory Background	60	48	33	21	18	11	11
Background Colour	8	21	9	6	6	13	2
Braille	3	5	2	6	2	2	1
Colour Overlay	28	46	43	16	11	7	1
Foreground Colour	8	21	9	6	6	13	2
General	32,427	30,660	24,066	29,502	32,674	18,238	33,464
Invert Colour Choice	6	8	9	2	10	4	0
Item-Translation Display	2	7	11	21	24	21	50
Keyword Translation Display	3	8	5	6	11	7	48
Large-Print Booklet	10	16	7	7	8	6	2
Masking	16	54	22	30	55	5	3
Onscreen Keyboard	28	19	14	7	11	14	0
Paper and Pencil	22	20	10	15	24	15	3
Signing	13	10	1	1	9	11	14
Spoken	3,668	3,944	3,128	3,055	3,213	1,662	1,450
Tactile	1	3	0	0	0	2	0
Total N	36,213	34,719	27,304	32,632	36,003	19,961	35,008

Part V

Technical Quality: Other

13
Full Performance Continuum

KAP WAS DEVELOPED to ensure that each subject-area and grade-level test provided reasonably precise estimates of student performance across the full performance continuum (i.e., from low-achieving to high-achieving students). This chapter provides evidence that supports this assertion.

13.1 Item Difficulty

TO HELP MEET THE GOAL stated above, KAP tests were developed to have a wide range of item difficulties. Summary results for KAP item difficulty are provided in both the classical item statistics chapter and the IRT item calibration chapter. The evidence presented in both of those chapters supports the conclusion that a very wide range of item difficulties were used on all ELA and math tests. For example, in terms of the classical item statistics, p-values values ranging from the single digits to high 0.90s were common.

13.2 Test Information/CSEMs

CSEMs (conditional standard errors of measurement) are documented visually in the reliability chapter. As noted in that chapter, the CSEM values increase at the extremes where scaled scores become very low and very high, resulting in the typical U-shaped pattern associated with IRT CSEMs. However, for many CSEM plots, the values change slowly across a fairly large range of scaled scores in the middle of the distribution, creating somewhat flat bottoms for the CSEM curves.

13.3 Cognitive Complexity

A related issue concerns the cognitive complexity of the KAP items. KAP items were categorized by cognitive complexity, as described by Webb.[1] Webb's DOK categories are:

[1] Webb (1997)

- *Level 1* (recall): requires simple recall of such information as a fact, definition, term, or simple procedure.
- *Level 2* (skill/concept): involves some mental skills, concepts, or processing beyond a habitual response; students must make some decisions about how to approach a problem or activity. Keywords distinguishing a Level 2 item include classify, organize, estimate, collect data, and compare data.
- *Level 3* (strategic thinking): requires reasoning, planning, using evidence, and thinking at a higher level.
- *Level 4* (extended thinking): requires complex reasoning, planning, developing, and thinking, most likely over an extended time. Cognitive demands are high, and students are required to make connections both within and among subject domains.

The DOK associated with each Target identifies the maximum DOK for an item. Items at Level 4, extended thinking, are not typically seen in most assessments unless performance tasks are included.

Items were developed to have a range of difficulties to support four performance levels as well as the eventual development of a stage-adaptive test.

Item complexity will be affected by the familiarity of the constructs being measured. Constructs that were previously taught in the same grade or earlier than described by the KCCRS will likely tend to appear easier in the early years of the assessment than constructs that were previously taught in higher grades or not addressed in previous content standards.

In the future, the final stage of the stage-adaptive test may include items that measure the content expectations from the next grade. This would be done to sufficiently challenge the most accomplished students, and to fairly claim that in fact some students exceed the content expectations at their chronological grade level.

The DOK results are summarized in the tables in the page margin. Most math items were at Level 2 and a few were at Level 3. In ELA most items were at Level 2 as well, but there were relatively more Level 3 items. There was only one Level 4 item, but as noted earlier, these are very rare in most assessments unless extended performance tasks are included.

Table 13.1: Item Counts by DOK Level and Grade for Math

Grade	Lvl 1	Lvl 2	Lvl 3
3	103	113	4
4	95	135	6
5	96	121	1
6	75	111	1
7	86	109	2
8	73	113	8
10	60	125	4

Table 13.2: Item Counts by DOK Level and Grade for ELA

Grade	Lvl 1	Lvl 2	Lvl 3	Lvl 4
3	83	194	34	0
4	70	174	44	0
5	86	160	43	0
6	76	164	45	0
7	53	177	38	1
8	60	152	38	0
10	61	199	31	0

14
Fairness and Accessibility

According to the *Standards for Educational and Psychological Testing*:

> The central idea of fairness in testing is to identify and remove construct-irrelevant barriers to maximal performance for any examinee.[1]

[1] AERA, APA, & NCME, 2014, p. 74

This language clearly identifies fairness as an issue related to the validity of test score inferences. Evidence in support of any assertion about the fairness of an assessment can come from several sources. This chapter addresses some of these evidence sources for the KAP.

14.1 Item and Test Development

Universally design (UD) in item and test development not only allows for the participation of the widest possible range of students, but it should also bolster the validity of score inferences as well. KAP's comprehensive inclusion rules mean that for all intents and purposes KAP tests include virtually all Kansas students. While initially motivated to meet the interests of special-needs students, the benefits of universally designed assessments should apply to all students with diverse characteristics.

See the chapter on inclusion, demographics, and participation for additional information.

During item writer training for KAP, participants received instruction on UD concepts. Item writer training included a definition of UD and examples of test items that adhered to UD principles. Additionally, the item writer guidelines included many UD principles. While details are provided elsewhere in this technical manual, the follow will give the reader some high-level examples of UD in the KAP's development:

See the chapters on item and test development for additional information about item writing and item review processes.

- Item writer training included an awareness of, and sensitivity to, issues of cultural and regional diversity.

- Both internal and external reviewers of items and test specifications ensured that no barriers were created due lack of sensitivity with respect to disability, culture, or other subgroups.

- The tests were developed to be compatible with many accommodations and a variety of widely used adaptive equipment and assistive technology.

- The language used on test materials was direct and concise. Additionally, unnecessary images and text were omitted as they were a potential distraction for students.

Ideally, accommodations should not change the meaning or difficulty of test items. Unfortunately, sample sizes are often too small to research this empirically.

14.2 Inclusion and Accommodations

Using appropriate item and test development processes is an excellent start for helping ensure fairness. As suggested above, even with the right processes, some students will need remaining barriers addressed. Test inclusion and accommodations policies help addresses these needs. This technical manual provides separate chapters for both inclusion and accommodations. The almost universal use of computers in KAP assessments actually facilitates many accommodations as many student needs can be meet through this administration format (magnification, text-to-speech, image contrasts, etc.) Some students must have paper or Braille tests, and these are made available to those students who needed them.

14.3 Differential Item Functioning

Differential item functioning (DIF) is statistical in nature and occurs when examinees with the same ability level but different group memberships do not have the same probability of answering an item correctly. This pattern of results may suggest the presence of item bias. As a statistical concept, DIF can be differentiated from general item sensitivity and bias issues, which are often content related and arise when an item presents negative group stereotypes, uses language that is more familiar to one subpopulation than to another, or is presented in a format that disadvantages certain learning styles. While the source of some sensitivity and bias issues are often plain to trained item reviewers, DIF may have no clear cause at times. Studying how DIF arises in an assessment program like KAP can provide information about how to better detect it and correct for it.

14.3.1 Some Limitations of Statistical Detection

No statistical procedure should be used as a substitute for rigorous, hands-on reviews by content and bias specialists. The statistical results can help organize the review so the effort is concentrated on the most suspicious cases. Further, no items should be automatically rejected simply because a statistical method flagged them or accepted because they were not flagged.

Statistical detection of DIF is not an exact science. There have been a variety of methods proposed for detecting DIF, but no single statistic can be considered either necessary or sufficient. Different methods are more or less successful depending on the situation. No analysis can guarantee that a test is free of bias, but almost any thoughtful analysis will uncover the most flagrant problems.

A fundamental shortcoming of all statistical methods used in DIF evaluation is that all are intrinsic to the test being evaluated. If a test is unbiased overall but contains one or two DIF items, any method will locate the problems. If, however, all items on the test show consistent DIF to the disadvantage of a given subpopulation, a statistical analysis of the items will not be able to separate DIF effects from true differences in achievement.

14.3.2 Logistic Regression Procedure for DIF

Logistic regression (LR) was used to test for DIF in KAP items. The approach is to predict one binary or ordinal item response from a summed score for the test, the grouping variable, and possibly the interaction between test score and group membership (Swaminathan & Rogers, 1990). In these analyses, conditioning was done using the total test score (for each subject and grade), but only uniform DIF was investigated (i.e., the interaction was not included in the LR models). The inclusion of the summed score controls for ability, and the uniform DIF test is statistically significant if the grouping variable is significant ($\alpha < .05$). Two group membership pairs were studied: male/female and African American/White.

With a large sample size, many trivial effects are statistically significant, so the practical significance was evaluated using the change in the Nagelkerke R^2 (between the models with and without the grouping variable) as an effect size to decide whether an item should be flagged as functioning differently between groups. The change in R^2 provides an indication of how much variance in the item response can be explained by group membership (e.g., gender). According to the recommendations of Jodoin and Gierl (2001), we considered values of R^2 less than .035 to be negligible levels of DIF, values between .035 and .07 to be moderate, and values above .07 to be large.

14.3.3 Results and Observations

For each subject, grade, and grouping variable where DIF analysis was conducted, counts of the number of items that showed statistically significant DIF are provided. Although many items showed statistical significance according to the ($\alpha < .05$) criterion, the sample sizes for KAP were fairly large and provided considerable statistical power in terms of the ability to reject the null hypothesis regarding group differences. When practical significance was evaluated, all test items had negligible effect sizes. The maximum change in the Nagelkerke R^2 was only 0.017 over all ELA items and 0.018 over all math items.

Grade	Items	Stat. Sig. Items	Favor Female	Favor Male
3	220	117	68	49
4	238	135	70	65
5	218	103	50	53
6	187	113	57	56
7	198	106	52	54
8	194	102	50	52
10	189	100	58	42

Table 14.1: Math Gender DIF: Item Counts for Statistical Significance by Grade

Grade	Items	Stat. Sig. Items	Favor African Am.	Favor White
3	220	51	19	32
4	238	61	22	39
5	218	45	16	29
6	187	45	15	30
7	198	54	18	36
8	194	41	15	26
10	189	35	11	24

Table 14.2: Math Race DIF: Item Counts for Statistical Significance by Grade

Grade	Items	Stat. Sig. Items	Favor Female	Favor Male
3	311	120	66	54
4	288	128	67	61
5	289	163	100	63
6	285	143	80	63
7	269	140	74	66
8	250	134	74	60
10	291	161	99	62

Table 14.3: ELA Gender DIF: Item Counts for Statistical Significance by Grade

Grade	Items	Stat. Sig. Items	Favor African Am.	Favor White
3	311	59	24	35
4	288	59	19	40
5	289	58	27	31
6	285	51	17	34
7	269	54	18	36
8	250	53	16	37
10	291	47	18	29

Table 14.4: ELA Race DIF: Item Counts for Statistical Significance by Grade

15
Performance Scoring

In future years the KAP ELA assessment will include multidiminsional performance tasks (MDPTs). The MDPTs will require human readers to evaluate the student responses to those tasks and score them against an appropriate rubric. This chapter is included now solely as a placeholder for future technical manuals. In the future, this chapter will include information about the following:

- Rangefinding (e.g., scoring leadership discussion of the rubric scoring guideline using representative samples of student responses at each score point)
- Rater recruitment/qualifications
- Leadership recruitment/qualifications
- training
- Handscoring process
- Handscoring validity process (e.g., validity papers used to check scorer accuracy)
- Quality control in scoring (e.g. reports used to monitor scorers)
- Descriptive statistics (e.g., score point frequencies)

Rater agreement proportions (adjacent and exact) and other indicators of scorer consistency will be included in future reliability chapters.

16
Classical Item Statistics

This chapter summarizes the classical item analysis results—i.e., item difficulty (how easy or hard an item is) and item discrimination (how much higher mean test scores are for students who answer an item correctly compared to those for students who answer the same item incorrectly)—for the KAP. These statistics represent the item characteristics most often used to determine (1) whether an item functioned properly during test administration and (2) how a group of students performed on an item.

16.1 Review of KAP Item Types

The KAP item types are listed below. As seen in the tables on the right, math item scores range from zero to two, and ELA item scores range from zero to three. MC-K items are worth one point. Other item types are not always associated with specific score values; content is the primary consideration in assigning point values to non-MC-K items. The KAP item types include

- selected-response (SR) items
  - multiple-choice keyed (MC-K) items: *selection items with a single correct response*
  - multiple-choice multiple-select (MC-MS) items: *selection items with (potentially) multiple correct responses*
- constructed-response (CR) items
  - short constructed-response (SCR) items: *shorter free-response items typically worth one point*
  - extended constructed-response (ECR) items: *longer free-response items typically greater than one point*

Table 16.1: Item Counts by Item Type and Grade for Math

Grade	CR	ITP	MC-K	MC-MS
3	19	37	144	20
4	10	58	141	29
5	15	40	141	22
6	11	33	130	13
7	13	31	135	19
8	9	32	137	16
10	2	21	148	18

Table 16.2: Item Counts by Item Points and Grade for Math

Grade	1-Point	2-Point
3	219	1
4	233	5
5	216	2
6	187	0
7	198	0
8	192	2
10	188	1

Table 16.3: Item Counts by Item Type and Grade for ELA

Grade	ITP	MC-K	MC-MS
3	78	209	24
4	65	194	29
5	51	203	35
6	60	198	27
7	66	173	30
8	52	177	21
10	73	188	30

Table 16.4: Item Counts by Item Points and Grade for ELA

Grade	1-Point	2-Point	3-Point
3	278	33	0
4	243	45	0
5	237	52	0
6	248	37	0
7	211	50	8
8	210	39	1
10	236	48	7

- innovative task package (ITP) items
  - background-graphic items
  - drop-down menu items
  - labeling items
  - matching-line items
  - matrix-interaction items
  - multiple-drop bucket items
  - ordering items
  - partition items
  - selection items
  - plotting-point items
  - straight line items
  - Venn-diagram items

ITP items are also known as technology enhanced items (TEIs) The last three ITP task types are currently used in math only. Examples of the ITPs used on the KAP assessment may be viewed through the KAP practice tests. Instructions for doing this are provided at: http://www.ksassessments.org/practice-tests.

16.2 Item-Level Statistics

16.2.1 Item Difficulty

ITEM DIFFICULTY is an important consideration for the KAP tests because of the range in performance levels defined in the state (i.e., students can fall into four performance levels). On standards-referenced tests like the KAP, test development aims to include a wide range of item difficulties. Items that are either extremely hard or extremely easy provide little information about differences in student achievement.

An item's *difficulty* is indicated by its mean score in a specified group (e.g., grade level). To get an item's mean score using the formula to the right, sum the individual item scores (x_i), and then divide by the total number of students who responded to the item (n). For dichotomously scored items (frequently used with the MC-K item type), student scores are represented by 0 (wrong) or 1 (right). With 0 – 1 scoring, the equation also represents the proportion of students correctly answering the item and the resulting statistic is called the item's p-value.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Figure 16.1: Mean item score formula.

In theory, p-values can range from 0.00 to 1.00 on the proportion-correct scale. For example, if an item has a p-value of 0.88, 88 percent of the students answered the item correctly. For CR items, mean scores can range from the minimum possible score—usually zero—to the maximum possible score (e.g., four points in the case of some math items). It is rare to see the minimum and maximum extremes of the difficulty scale in applied practice. However, understanding the

Pseudo p-values can be provided for CR items. This is done by dividing the mean item score by the maximum possible item score.

extremes helps illustrate that relatively lower values correspond to more difficult items and relatively higher values to easier items.

As a specific example of how knowledge of an item's minimum and maximum score values can assist with interpretation, consider that an item p-value of 0.93 out of a maximum p-value of 1.0 suggests either the item is relatively easy or the students who attempted the item were relatively high achievers. In other words, item difficulty and student academic achievement are often difficult to differentiate.

With this caveat in mind, an item answered correctly by a high percentage of students might suggest most students have mastered the knowledge or skill the item taps. Conversely, an item answered correctly by a low percentage of students might suggest few students have mastered the knowledge or skill the item taps.

For difficult MC-K items with four response options, complete random guessing by students would lead to an expected p-value of 1/4 (0.25). For MC-K items with five response options, the guessing p-value would be 1/5 (0.20), and so on for other numbers of response options.

16.2.2 Item Discrimination

Item discrimination is important for KAP because the use of discriminating items on a test is associated with reliable test scores. This in turn leads to (1) precise score estimates (i.e., there will be smaller confidence intervals around the scores) and (2) more accurate performance level placements.

Item *discrimination* reflects an item's ability to differentiate between those with high and those with low test scores. Ideally, high-achieving students (i.e., those who perform well on KAP overall) should be more likely to answer any given KAP item correctly whereas low-achieving students (i.e., those who perform poorly on KAP overall) should be more likely to answer the same item incorrectly.

For the KAP tests, the Pearson's product-moment correlation coefficient between student item scores and test scores indicates an item's discrimination. These are usually called *item-test correlations*, or when items have dichotomous (0,1) scores, *point-biserial correlations*.

One can interpret the point-biserial correlation as a standardized mean difference. A positive value indicates that students who got the item correct had a higher mean test score whereas a negative value indicates that students who got the item correct had a lower mean test score.

The correlation coefficient can range from –1.0 to +1.0. For items that tend to have more higher-scoring students answering correctly and more lower-scoring students answering incorrectly, the correlation between that item's scores and the students' total test scores will be both positive in value and moderately large in magnitude (i.e., well above zero). An item with such a correlation is a good discriminator between high- and low-ability students.

16.3 Summary of Item Statistics

When interpreting item difficulty and discrimination indices, one should consider whether an item is polytomously or dichotomously scored and what the maximum possible point value is for the items.

16.3.1 Mathematics

16.3.1.1 Difficulty

Descriptive statistics for the item difficulty of the one-point items in math are tabulated below. The median (or P_{50}) p-values for these items ranged from 0.40 to 0.55. Based on the median difficulty for these items, they appear challenging for most students. A very wide range of item difficulties exists which spans nearly the entire range of possible p-values. Having a wide range of item difficulties on each grade-level exam was one test development goal. Fewer than one half of one percent of students answered the most difficult item correctly whereas 97% of students answered the easiest item.

Because so few items were worth two points, summary breakouts are only provided for one-point items.

The associated figure illustrates how item difficulties are dispersed by item type. This figure includes both one- and two-point items; however, the two-point items were converted to pseudo p-values so they are on the same $0 - 1$ scale as the one-point items. The hardest items at most grade levels were generally non-MC-K items, especially at the upper grades where the hardest items were CR or ITP items.

Grade	k	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
3	220	0.09	0.24	0.37	0.53	0.69	0.84	0.96
4	238	0.05	0.24	0.35	0.55	0.68	0.81	0.95
5	218	0.02	0.21	0.31	0.48	0.62	0.77	0.88
6	187	0.12	0.26	0.34	0.47	0.61	0.74	0.97
7	198	0.01	0.19	0.28	0.40	0.57	0.67	0.90
8	194	0.02	0.16	0.27	0.41	0.54	0.67	0.86
10	189	0.00	0.20	0.30	0.42	0.53	0.62	0.87

Table 16.5: Item Difficulty Summary Statistics for Math One-Point Items

k = the number of test items.

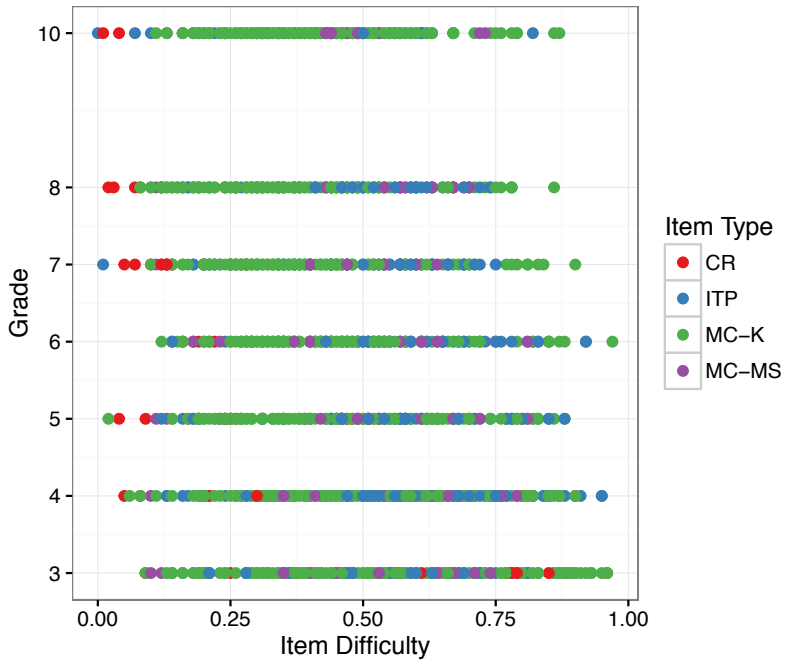


Figure 16.2: Math Item Difficulty Dot Plot by Grade and Item Type

16.3.1.2 Discrimination

Descriptive statistics for the item discrimination of the dichotomous and polytomous items in math are tabulated below. The median item-total correlations for dichotomous items ranged from about 0.33 to 0.41. For polytomous items, the range was 0.33 to 0.45.

The item-test correlations presented here were *not* corrected for spuriousness (i.e., the item scores were not removed from the total scores before the correlations were computed). As a result, these values will be slightly larger than some might expect.

The figure illustrates how item-test correlations are dispersed by the number of response categories. Generally, there was a good degree of overlap in item difficulty across the response categories. However, items with dichotomous response categories often had the lowest item-total correlations. This result is understandable as range restriction has a well-known effect on correlations.

Grade	k	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
3	220	0.14	0.24	0.31	0.39	0.48	0.53	0.59
4	238	0.11	0.25	0.32	0.39	0.46	0.53	0.63
5	218	0.15	0.25	0.32	0.41	0.48	0.52	0.67
6	187	0.11	0.21	0.28	0.39	0.45	0.53	0.62
7	198	0.07	0.21	0.27	0.36	0.43	0.49	0.61
8	194	0.10	0.22	0.27	0.36	0.42	0.47	0.56
10	189	0.01	0.21	0.25	0.33	0.41	0.49	0.62

Table 16.6: Item Discrimination Summary Statistics for Math Dichotomous Items

At Grade 10, the minimum item-test correlation was 0.01. Although low, this item was approved by content experts who found no flaws with this item. This item's IRT a-parameter estimate was > 0.30.

Grade	k	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
3	220	0.21	0.26	0.33	0.38	0.45	0.50	0.52
4	238	0.20	0.27	0.38	0.45	0.50	0.58	0.66
5	218	0.18	0.29	0.33	0.39	0.46	0.54	0.68
6	187	0.14	0.21	0.25	0.33	0.41	0.43	0.49
7	198	0.18	0.24	0.28	0.35	0.40	0.45	0.62
8	194	0.19	0.23	0.29	0.35	0.43	0.49	0.55
10	189	0.13	0.18	0.26	0.35	0.39	0.47	0.55

Table 16.7: Item Discrimination Summary Statistics for Math Polytomous Items

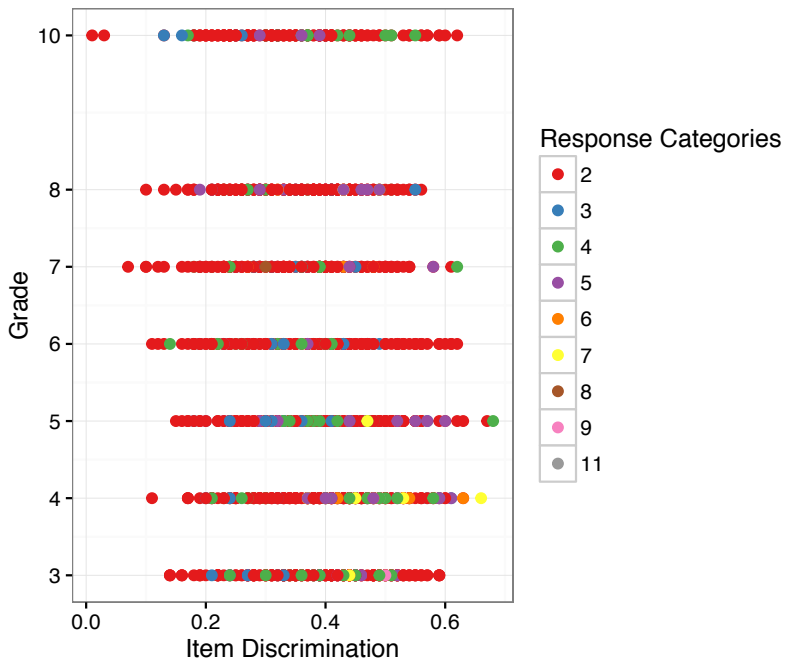


Figure 16.3: Math Item Discrimination Dot Plot by Grade and Number of Response Categories

16.3.2 ELA

16.3.2.1 Difficulty

THE FOLLOWING TABLES break-out the item difficulty results for KAP items by grade level. The median p-values ranged from about 0.53 to 0.66 for the one-point items and 1.12 to 1.36 for the two-point items. From the difficulty distributions illustrated in the plots, a wide range of item difficulties appeared on each exam, which met one test development goal.

The associated figure illustrates how item difficulties are dispersed by item type. This figure includes all items; however, the two- and three-point items were converted to pseudo p-values so they are on the same $0 - 1$ scale as the one-point items. The hardest items at most grade levels were ITP items; however, this item type also appeared among the easier items as well.

Grade	k	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
3	311	0.11	0.32	0.41	0.53	0.65	0.75	0.90
4	288	0.07	0.40	0.50	0.61	0.74	0.81	0.95
5	289	0.14	0.35	0.54	0.66	0.79	0.85	0.95
6	285	0.17	0.35	0.45	0.60	0.74	0.82	0.95
7	269	0.16	0.40	0.47	0.61	0.72	0.80	0.93
8	250	0.06	0.36	0.46	0.60	0.71	0.82	0.90
10	291	0.01	0.34	0.45	0.56	0.72	0.80	0.92

Table 16.8: Item Difficulty Summary Statistics for ELA One-Point Items

Grade	k	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
3	311	0.20	0.46	0.92	1.14	1.29	1.47	1.77
4	288	0.30	0.57	0.83	1.18	1.47	1.54	1.80
5	289	0.26	0.69	1.07	1.31	1.50	1.67	1.75
6	285	0.41	0.85	1.10	1.28	1.41	1.52	1.63
7	269	0.25	0.41	0.67	1.12	1.48	1.62	1.79
8	250	0.34	0.58	1.03	1.36	1.53	1.65	1.77
10	291	0.14	0.59	0.92	1.18	1.44	1.59	1.83

Table 16.9: Item Difficulty Summary Statistics for ELA Two-Point Items

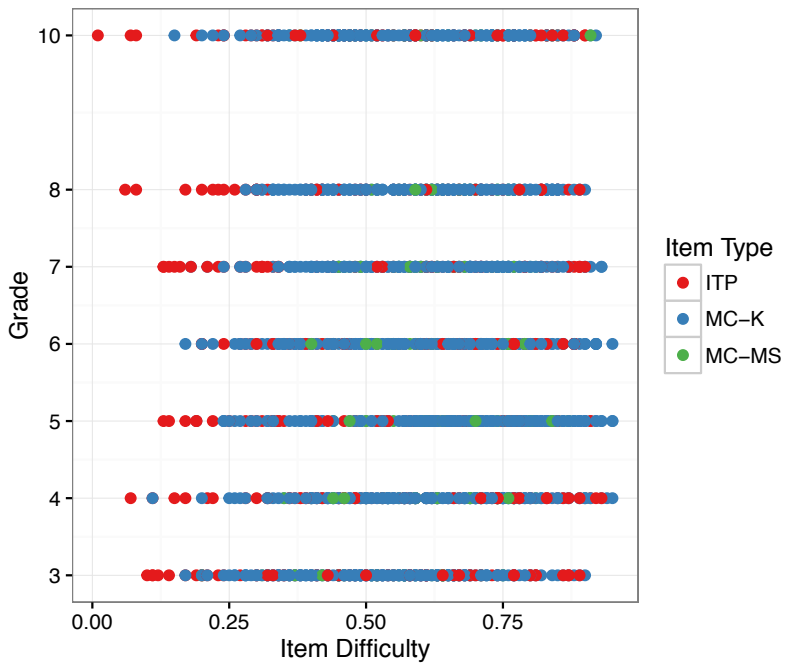


Figure 16.4: ELA Item Difficulty Dot Plot by Grade and Item Type

16.3.2.2 Discrimination

Descriptive statistics for the item discrimination of the dichotomous and polytomous items in ELA are tabulated below. The median item-total correlations for dichotomous items ranged from about 0.35 to 0.37. For polytomous items, the range was 0.42 to 0.47.

The associated figure illustrates how item-total correlations are dispersed by the number of response categories. Items with dichotomous categories tended to have lower item-total correlations. This was the case for math tests as well, although it is even more prominant in ELA tests.

Grade	k	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
3	311	0.15	0.21	0.29	0.37	0.44	0.49	0.62
4	288	0.13	0.23	0.29	0.36	0.42	0.47	0.66
5	289	0.13	0.23	0.29	0.36	0.42	0.48	0.59
6	285	0.15	0.23	0.29	0.36	0.43	0.47	0.58
7	269	0.14	0.21	0.27	0.35	0.41	0.46	0.60
8	250	0.14	0.24	0.29	0.35	0.41	0.47	0.63
10	291	0.13	0.22	0.28	0.35	0.42	0.46	0.55

Table 16.10: Item Discrimination Summary Statistics for ELA Dichotomous Items

Grade	k	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
3	311	0.20	0.31	0.38	0.44	0.51	0.57	0.59
4	288	0.20	0.33	0.37	0.44	0.51	0.56	0.64
5	289	0.23	0.32	0.39	0.44	0.51	0.56	0.64
6	285	0.15	0.26	0.36	0.42	0.49	0.53	0.65
7	269	0.19	0.28	0.36	0.43	0.50	0.56	0.60
8	250	0.21	0.27	0.38	0.44	0.50	0.55	0.65
10	291	0.23	0.30	0.37	0.47	0.53	0.55	0.60

Table 16.11: Item Discrimination Summary Statistics for ELA Polytomous Items

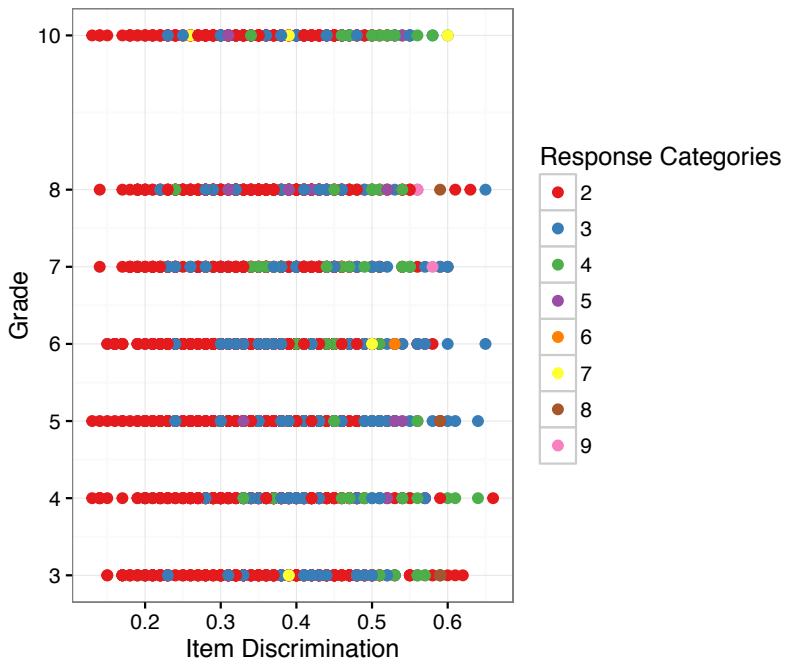


Figure 16.5: ELA Item Discrimination Dot Plot by Grade and Number of Response Categories

16.4 Additional Visualizations

NEXT, A SERIES OF SCATTER PLOTS show item discrimination values (y-axis) and item difficulty values (x-axis) for each grade and subject-area test. These plots provide a considerable amount of information about item discrimination and difficulty in a single visual image for each KAP test. This is due, in part, to the x- and y-axes including *rug plots* and a *seven-number* statistical summary:

- minimum and maximum values
- median
- four percentile values (P_{10}, P_{25}, P_{75}, and P_{90})

These results included all operational items.

No specific relationship is expected between the item discrimination values (i.e., item-test correlations) and item difficulty values (i.e., item mean scores). However, an interaction can exist between item discrimination and item difficulty. Items answered correctly or incorrectly by a large proportion of examinees (i.e., the items have extreme p-values) can have reduced power to discriminate, and thus, can have lower item-test correlations. These plots can illustrate cases where items might have lower discrimination values because of their extreme difficulties. In the following visuals, such items appeared more frequently with harder test items than easier test items.

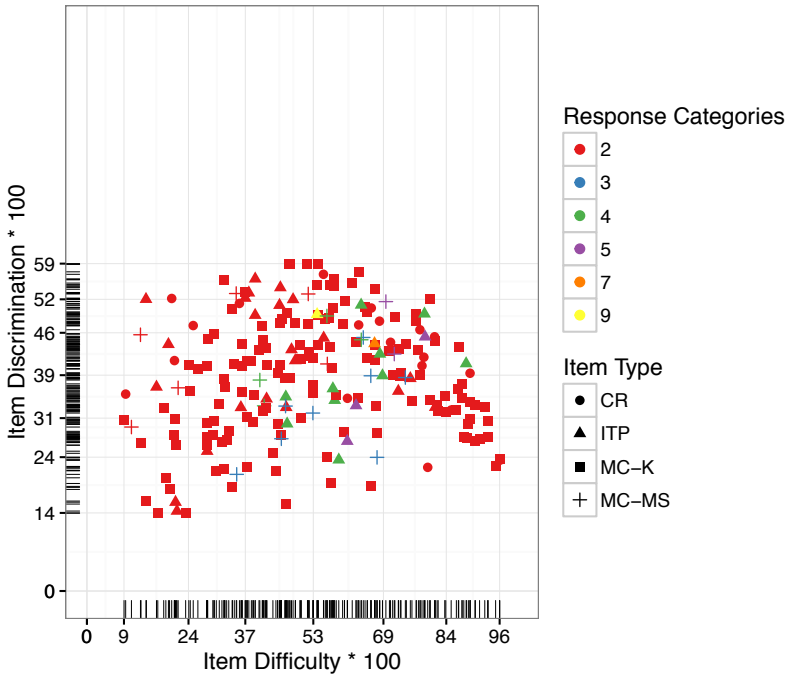


Figure 16.6: Item-Test Correlation on Item Difficulty: Grade 3 Math

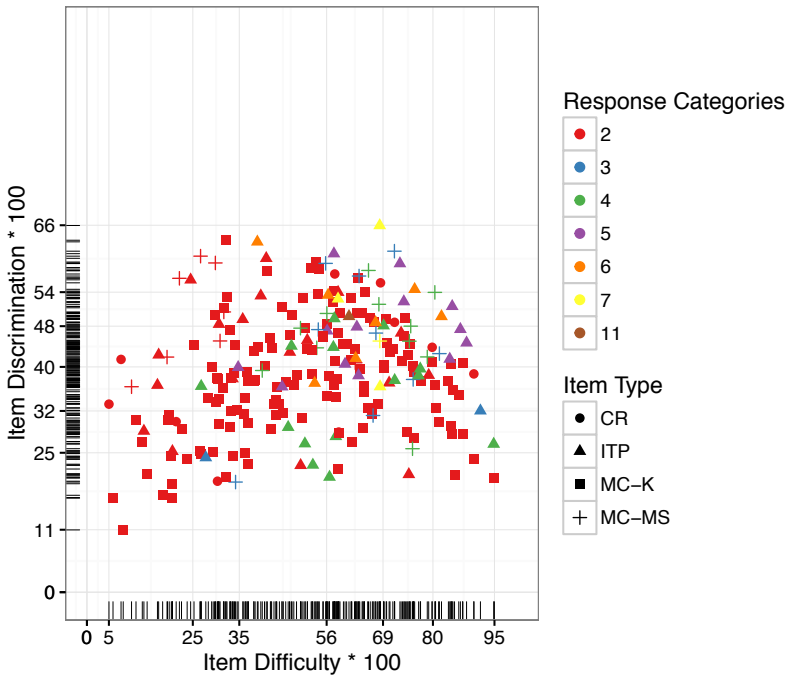


Figure 16.7: Item-Test Correlation on Item Difficulty: Grade 4 Math

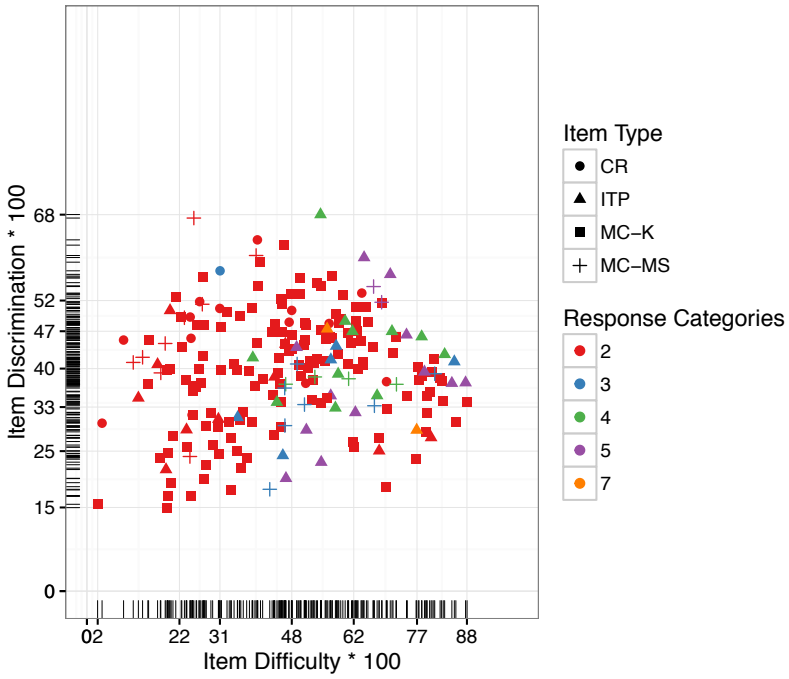


Figure 16.8: Item-Test Correlation on Item Difficulty: Grade 5 Math

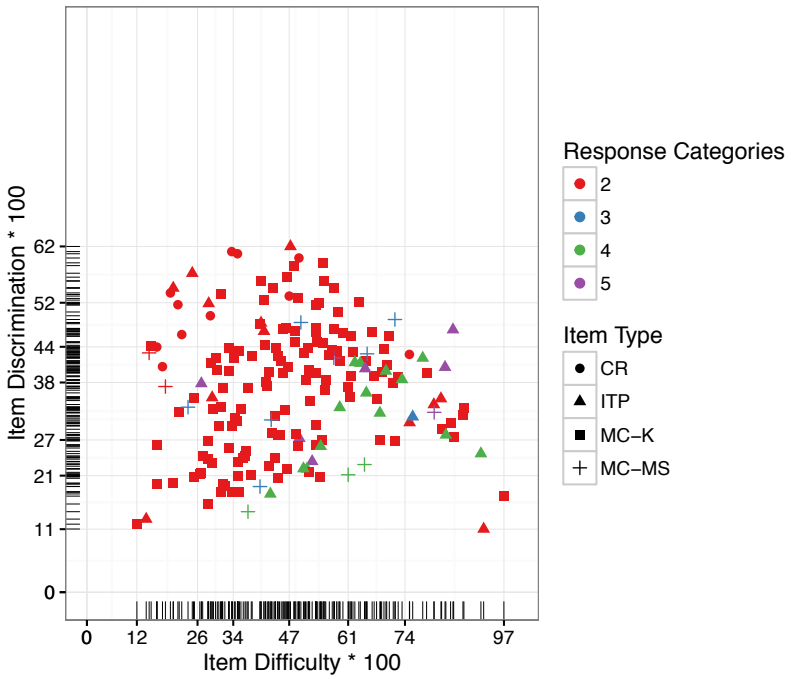


Figure 16.9: Item-Test Correlation on Item Difficulty: Grade 6 Math

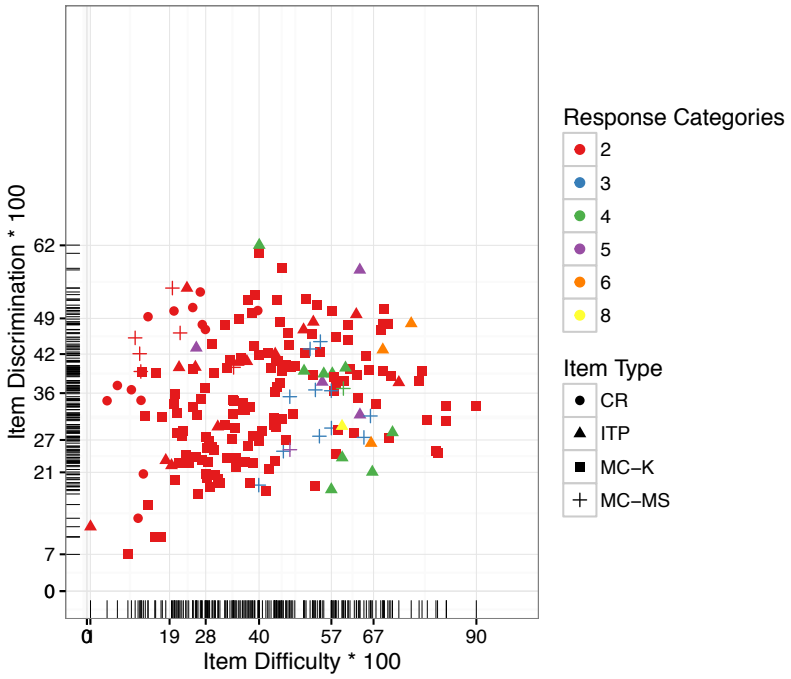


Figure 16.10: Item-Test Correlation on Item Difficulty: Grade 7 Math

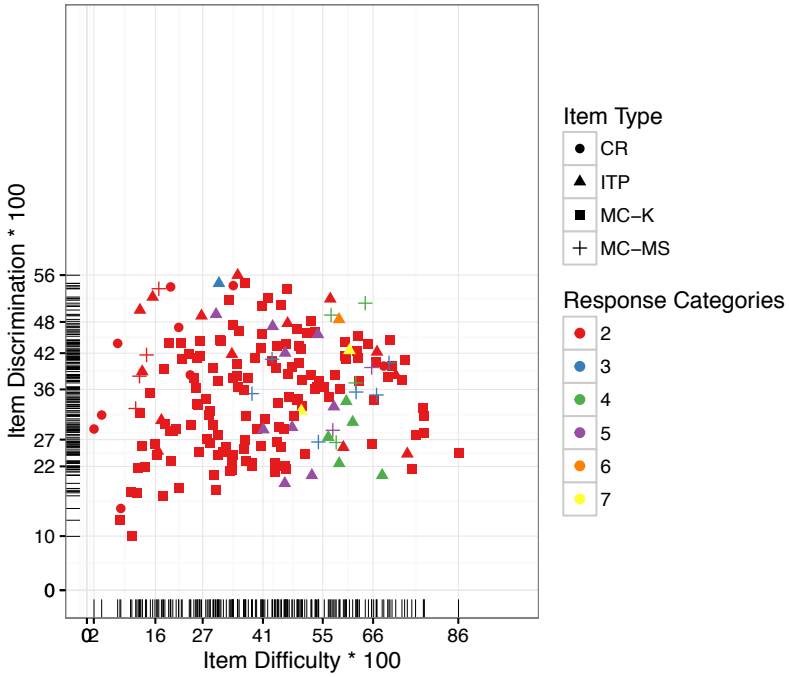


Figure 16.11: Item-Test Correlation on Item Difficulty: Grade 8 Math

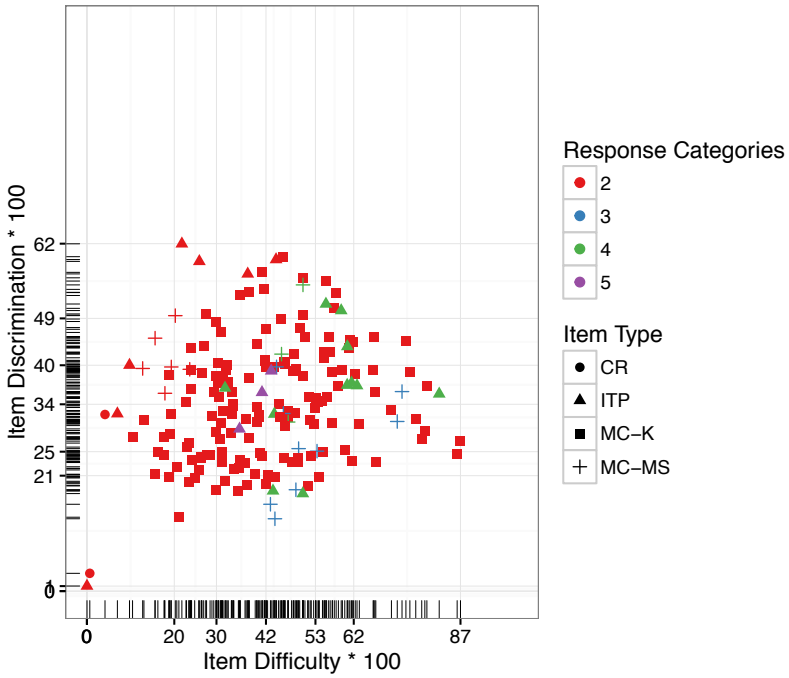


Figure 16.12: Item-Test Correlation on Item Difficulty: Grade 10 Math

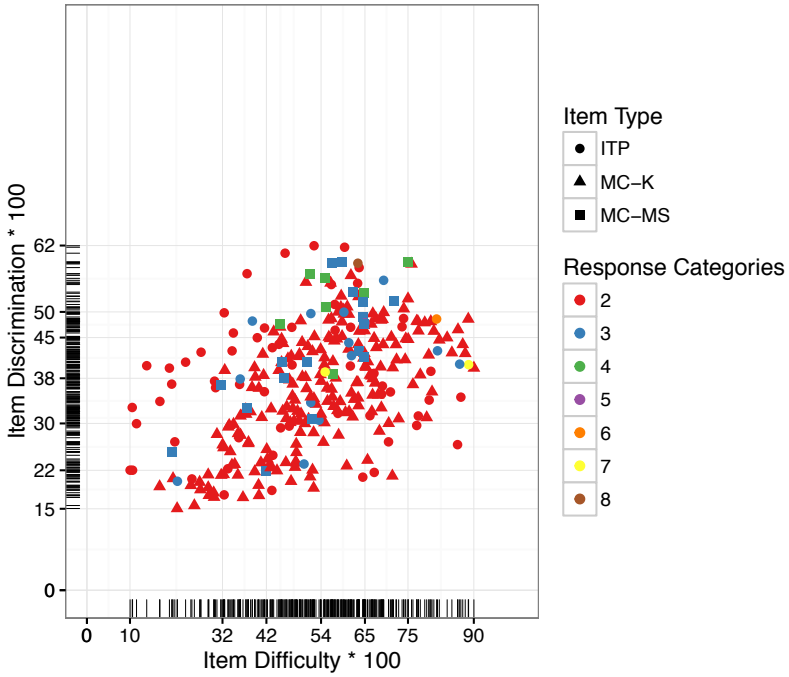


Figure 16.13: Item-Test Correlation on Item Difficulty: Grade 3 ELA

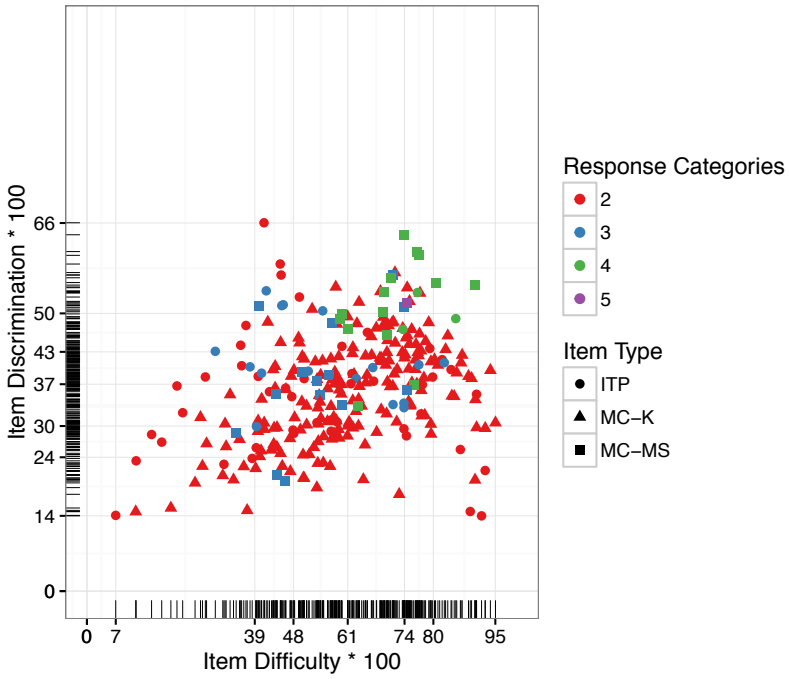


Figure 16.14: Item-Test Correlation on Item Difficulty: Grade 4 ELA

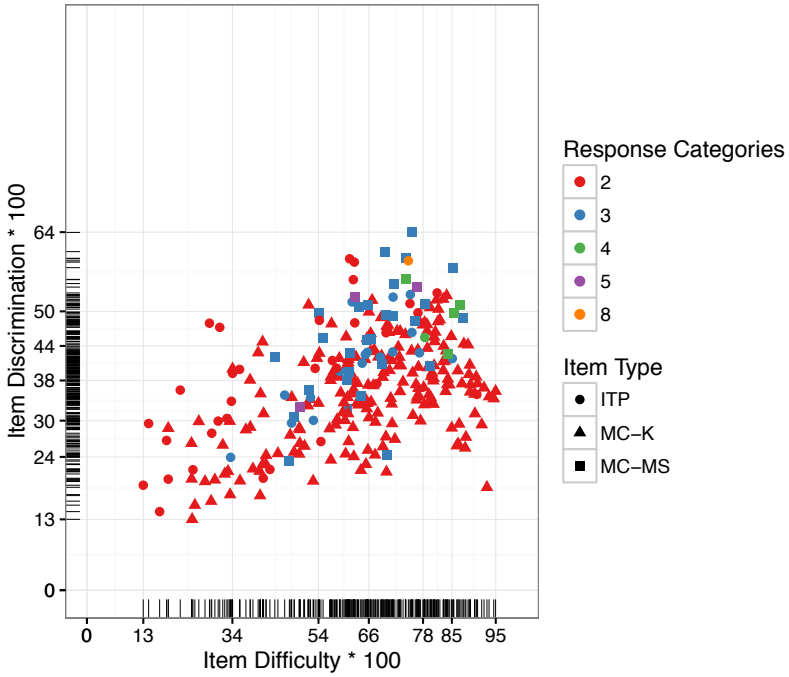


Figure 16.15: Item-Test Correlation on Item Difficulty: Grade 5 ELA

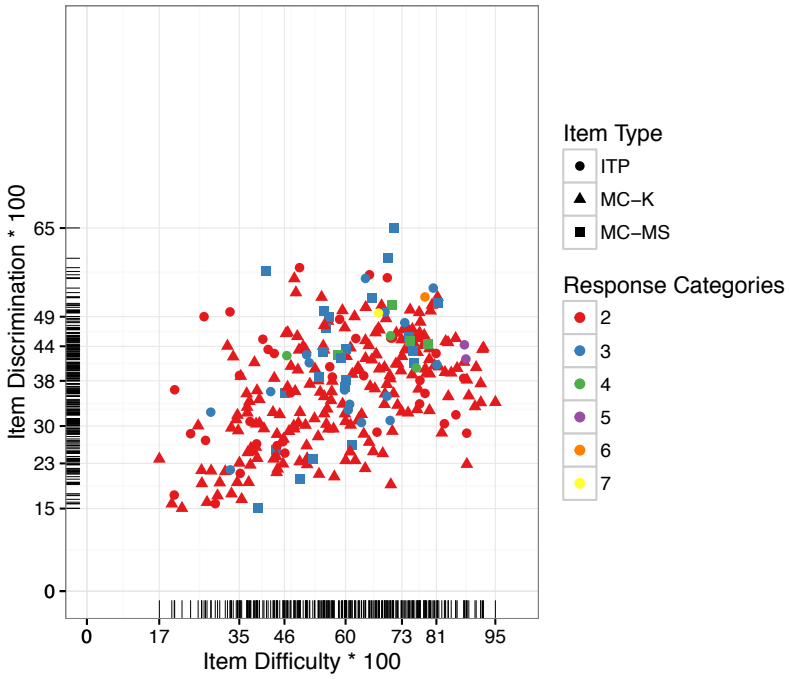


Figure 16.16: Item-Test Correlation on Item Difficulty: Grade 6 ELA

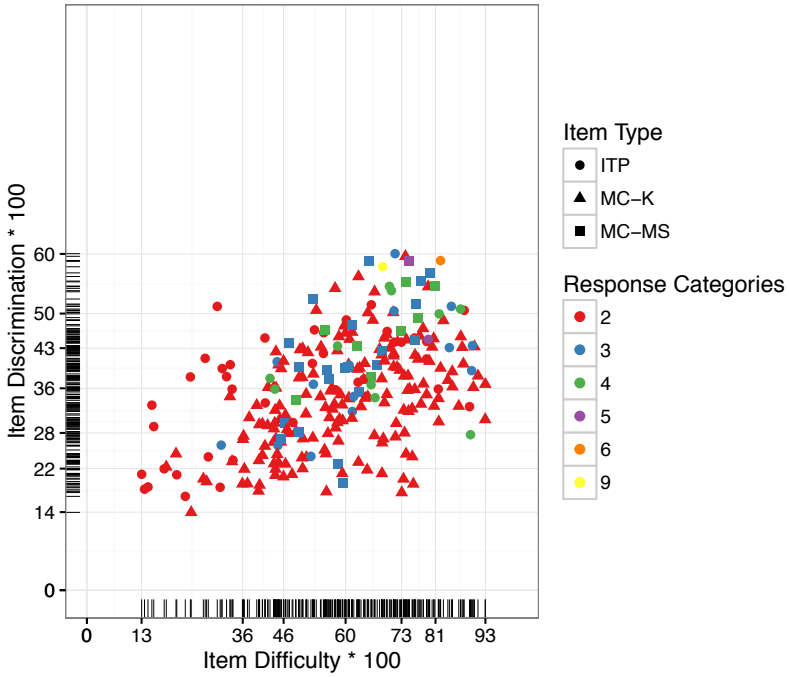


Figure 16.17: Item-Test Correlation on Item Difficulty: Grade 7 ELA

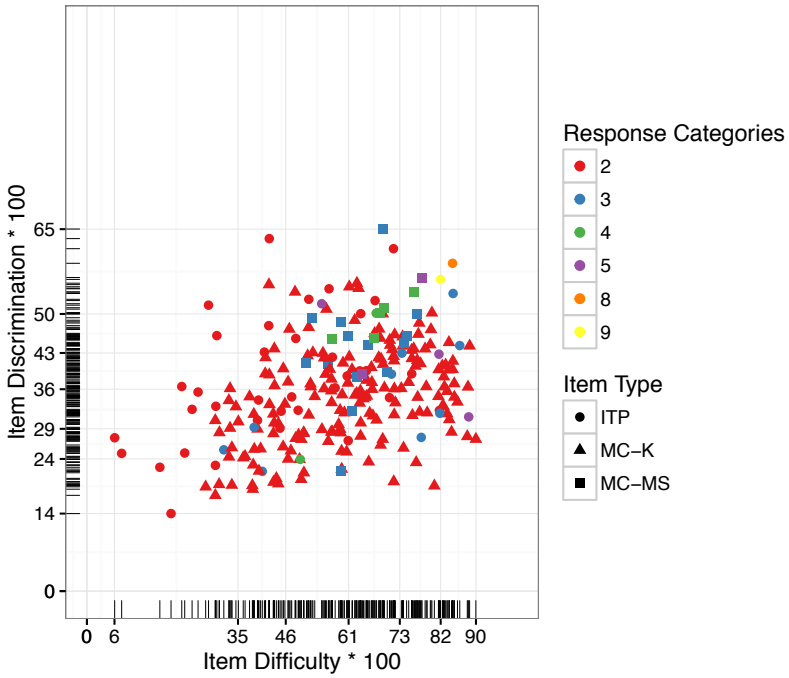


Figure 16.18: Item-Test Correlation on Item Difficulty: Grade 8 ELA

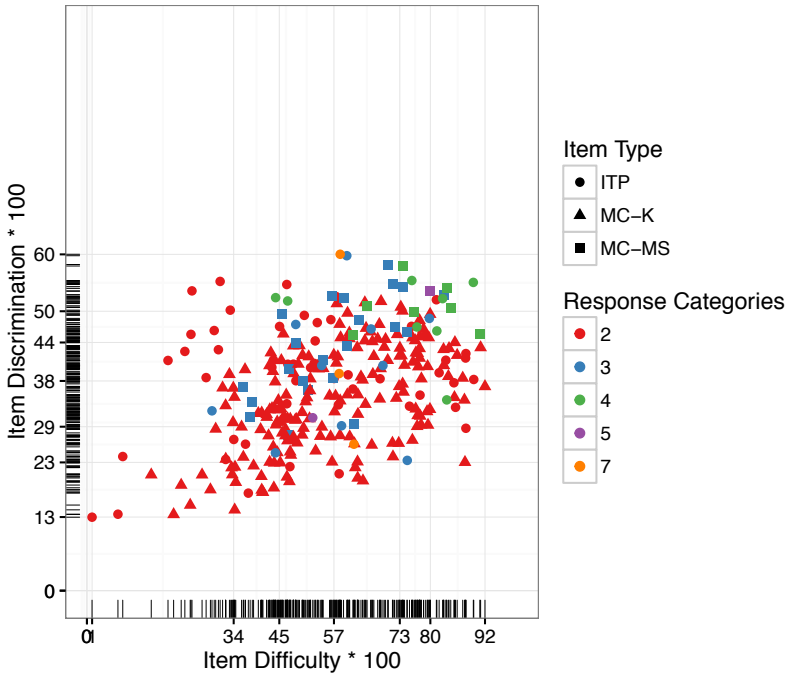


Figure 16.19: Item-Test Correlation on Item Difficulty: Grade 10 ELA

16.5 Summary

IT IS DIFFICULT to make global conclusions about overall test quality from these item statistics alone. However, the results presented suggest that many KAP items are challenging in their difficulty yet provide good discrimination. Further, a very wide range of item difficulties exist for each grade and subject test, which is appropriate given the KAP tests are used to classify students into four performance levels. Test development staff will continue monitor these item statistics in future administrations.

17
IRT Item Calibration

Responses to KAP items were analyzed using item response theory (IRT). Although IRT is nearly an industry standard for item analysis in large-scale K – 12 assessment programs, IRT models make several strong assumptions related to dimensionality, local independence, model-data fit, and item parameter invariance. The resulting inferences from any application of IRT depend on the degree to which the underlying assumptions are met.

This chapter outlines the procedures used for calibrating KAP operational items. Generally, item calibration is the process of estimating the parameters for each item on an assessment so that all items are placed on a common scale. This chapter (1) introduces the two parameter logistic (2PL) and graded response IRT models which are used for KAP, (2) reports the results from the evaluation of several IRT assumptions, and (3) summarizes item statistics for the KAP Math and English Language Arts (ELA) tests.

17.1 Description of the IRT Models

The two-parameter logistic (2PL) model (Birnbaum, 1968) and the graded-response model (GRM) (Samejima, 1969, 1997) were used to calibrate KAP items because both multiple-choice (binary scored) and constructed-response (ordinally scored) items were included in the KAP assessments. The GRM is appropriate for ordinal responses and the 2PL is a special case of the graded model for dichotomous (0, 1 scored) responses.

Under the 2PL, the probability that the response u to item i is correct is:

$$P(u_i = 1|\theta) = \frac{e^{[a_i(\theta - b_i)]}}{1 + e^{[a_i(\theta - b_i)]}}$$

where a_i is the discrimination parameter, b_i is the difficulty parameter,

and θ represents latent proficiency that varies randomly over students. Discrimination indicates how well the item distinguishes between students with higher versus lower levels of proficiency, and difficulty is the degree of item difficulty on the same scale as θ.

Under the GRM, the probability that u_i is equal to any observed response category v is equal to the probability that the response is v or higher, less the probability that the response is higher than v:

$$P(u_i = v|\theta) = \frac{1}{1 + e^{[-a_i(\theta - b_{i,v-1})]}} - \frac{1}{1 + e^{[-a_i(\theta - b_{i,v})]}}$$

Note that one discrimination parameter is estimated for each item. This parameter may be interpreted as the strength of association between the item and θ. For m response categories, there are $m - 1$ GRM b_i parameters. The b_i for category v is interpreted as the point on θ where the probability is 0.5 of responding in category v or higher.

17.2 Calibration Procedures

17.2.1 Software and Estimation Algorithm

Item calibration was implemented via flexMIRT 2.80 (Cai, 2013) with the default estimation criteria for the expectation-maximization (EM) estimation method.

17.2.2 Sample

The estimation sample include all student who complete at least on test section. All omits (no response) were scored as incorrect answers (coded as 0s) for calibration. Every grade had eight forms that were randomly assigned to students. However, all students who needed an accommodation took Form A. Across the two subjects (Math and English Language Arts (ELA)) and seven grade levels (Grade 3 - 8 and high school), there were a total of 112 test forms.

Only the students who were randomly assigned to the test forms were used during item calibration.

The first 25 items on each test form were the same for all students. The data format for the calibration is the sparse matrix format for the concurrent calibration. Thus, all items on eight forms can be estimated and put on the same scale at the same time.

17.3 Evaluating IRT Assumptions

The validity of inferences from the IRT results depends on the degree to which assumptions of the models were met and how well the models fit the data. In this section, assumptions about unidimensionality, local item independence, item fit, and item parameter invariance are evaluated.

17.3.1 Marginal Fit for Items

The marginal χ^2 fit statistic was used to evaluate the model fit for individual items. FlexMIRT (Cai, 2013) computes this statistic during item calibration. The marginal χ^2 fit statistic of one item follows the χ^2 distribution with degrees of freedom equal to the number of categories for that item minus 1. Using a significance level 0.05, the following tables include the number of items, the number of misfit items and the percent of misfit items.

Table 17.1: Math Misfit Results by Grade

Grade	k	k Misfit	Proportion
3	220	6	0.03
4	238	16	0.07
5	218	53	0.24
6	187	26	0.14
7	198	17	0.09
8	194	38	0.20
10	189	36	0.19

k = number of test items

Table 17.2: ELA Misfit Results by Grade

Grade	k	k Misfit	Proportion
3	311	0	0.00
4	288	7	0.02
5	289	35	0.12
6	285	4	0.01
7	269	1	0.00
8	250	42	0.17
10	291	0	0.00

k = number of test items

17.3.2 Local Independence

The G^2 local dependence (LD) statistic (Chen & Thissen, 1997) was used to evaluate the local item independence assumption. For each item pair, flexMIRT (Cai, 2013) computed this statistic during the item calibration. The flexMIRT manual (Houts & Cai, 2013), suggests that values over 3.0 indicates the presence of LD between item pairs. The following tables include the number of item pairs, the number of item pairs presenting LD and the percent of items pairs presenting LD.

Grade	Total Item Pairs	Locally Dependent Pairs	Proportion
3	9793	1917	0.20
4	10632	3514	0.33
5	10080	4232	0.42
6	8633	3222	0.37
7	8560	1906	0.22
8	7991	3393	0.42
10	7964	3194	0.40

Table 17.3: Count of Locally Dependent Math Item Pairs by Grade

Grade	Total Item Pairs	Locally Dependent Pairs	Proportion
3	13425	2848	0.21
4	12283	3102	0.25
5	11883	4319	0.36
6	11522	3087	0.27
7	11326	2752	0.24
8	10283	4687	0.46
10	12609	3559	0.28

Table 17.4: Count of Locally Dependent ELA Item Pairs by Grade

17.3.3 Unidimensionality

Both the 2PL and GRM assume that all the items scaled together measure a single dominant latent variable. Confirmatory factor analysis (CFA) was applied to every test (i.e., to every form within subject and grade) to evaluate whether a model with one dominant dimension fit the data reasonably well. CFA was carried out using tetrachoric/polychoric correlations for binary/ordinal item responses and robust weighted least-squares estimation with the lavaan R package.[1] The one-factor CFA model was considered to fit well if the comparative fit index (CFI) and Tucker Lewis Index (TLI) were .95 or greater and the Root Mean Square Error of Approximation (RMSEA) was .05 or smaller.

[1] Rosseel, 2012

For KAP Math, all examinees responded in the same category (incorrect) for one item on Grade 10 Form E (Item 56) and one item on Grade 10 Form F (Item 54). As a result, these items were omitted from the respective CFA model. Otherwise, all items were included in the models. Results showed that all CFA models for both Math and ELA fit well. Over all the tests, the CFI ranged from .96 to 1.0, TLI ranged from .96 to 1.0, and RMSEA ranged from .01 to .03. All the tests may be reasonably treated as unidimensional.

17.3.4 Invariance

For IRT calibration and scoring, it is assumed that the item parameter estimates are invariant up to a linear transformation for all examinees. To evaluate this assumption (1) examinees were randomly assigned to two equal-sized samples (Sample X and Sample Y), (2) IRT item parameters were re-estimated for each group, and (3) the similarity between the estimates for the two groups evaluated.

Since KAP used multiple test forms, one form was selected for each subject and grade. To evaluate the similarity between the two sets of item parameters, bivariate scatter plots were created and the associated the Pearson product-moment correlations calculated. Difficulty parameters were evaluated for groups of items with the same number of response categories. Because most items on the tests were binary, there were sometimes very small numbers of items with a certain response options (e.g., 9). Difficulty parameters were plotted and correlated if there were at least 10 items with the particular number of response options.

As noted elsewhere in this chapter, some item parameters were extreme. To avoid biasing the correlations with extreme outliers, any difficulty parameter estimate more extreme than 6 in absolute value was recoded to -6 or 6. Scatter plots and correlations are available in the appendix. The invariance assumption would appear to be met if the item parameters estimated for Samples X and Sample Y were highly correlated with one another.

For both Math and ELA, the relationships between item parameter estimates for samples X and Y were very strong and linear, with almost all Pearson correlations near 1. The smallest observed Pearson correlation for ELA was .95. The smallest observed Pearson correlation for Math was .85 for discrimination in Grade 10. All other correlations for Math were .93 or above. Overall, these results indicate strong support for the invariance assumption for KAP Math and ELA.

Invariance cannot be taken for granted, but often holds, particularly over random samples of examines as was the case here. Had the results been different, data from additional forms would have been analyzed. Future technical manuals will continue to investigate invariance and more rigorous tests will be undertaken. For example, next year's manual will study invariance under samples that differ in ability, perhaps using groups formed on the bases of gender or ethnicity.

17.4 IRT Item Statistics

IRT analyses were carried out using concurrent calibration for each grade so that all of the item parameter estimates are on the same scale and may be summarized together across forms. The following tables

Note that FlexMIRT's default is to scale results so that the underlying latent ability distribution has a mean = 0 and an SD = 1.)

summarize the discrimination and difficulty parameter estimates for operational items on each test. The number of items reflected in the summary is given for each grade and parameter estimate below. Most items were dichotomous, but some items had as many as 11 categories (thus, 10 b parameters). Parameters for all items, irrespective of the number of categories, are included together in the tables below. When there was only a single item contributing the parameter estimate for a grade, the estimate is given in the minimum and maximum columns.

Although item discrimination was not usually too far from 1.0 on average, it clearly varied over items, justifying the use of models that permit that parameter to vary over items, in contrast to the simpler Rasch-family models. The mean item discrimination declined as grade increased for Math, but remained close to .9 in all grades for ELA.

Some estimated difficulty parameters were very extreme (outside the range -4 to +4), but for practical purposes such estimates may be considered about -4 (if negative) or +4 (if positive). It is a feature of the estimation procedure that estimates near the boundaries sometimes approach infinite, especially when there is less mass available to estimate them (i.e., fewer students in that range of proficiency). Recall that difficulty parameters and proficiency are on the same scale. The means are of course, pulled in the direction of the extreme values. The minima and maxima for the b parameters indicate that the items included in the KAP assessments adequately covered the full performance continuum. Should extreme parameters continue to be observed in future assessments, the KAP TAC will be consulted about mitigation strategies (e.g., establishment of prior distributions on the difficulty parameters).

Table 17.5: Summary Statistics for IRT Parameters

Subject	Grade	k	Parm	Mean	SD	Median	Min	Max
ELA	3	311	a1	0.916	0.412	0.865	0.300	2.496
ELA	3	311	b1	-0.347	1.429	-0.421	-5.382	4.156
ELA	3	42	b2	0.273	1.651	0.701	-2.130	2.433
ELA	3	12	b3	0.273	1.651	0.701	-2.130	2.433
ELA	3	5	b4	-0.592	1.146	-1.085	-1.914	0.755
ELA	3	4	b5	-0.257	1.145	-0.304	-1.609	1.191
ELA	3	3	b6	1.168	2.647	0.958	-1.367	3.914
ELA	3	1	b7				2.852	2.852
ELA	4	288	a1	0.918	0.369	0.867	0.306	2.147
ELA	4	288	b1	-0.919	1.503	-0.878	-5.686	5.171
ELA	4	48	b2	0.050	1.175	-0.296	-1.216	2.086

Continued on next page

Subject	Grade	k	Parm	Mean	SD	Median	Min	Max
ELA	4	18	b3	0.050	1.175	-0.296	-1.216	2.086
ELA	4	1	b4				0.864	0.864
ELA	5	289	a1	0.948	0.375	0.910	0.301	2.427
ELA	5	289	b1	-1.097	1.598	-1.240	-9.624	5.233
ELA	5	54	b2	-0.617	0.882	-0.667	-2.182	0.893
ELA	5	9	b3	-0.617	0.882	-0.667	-2.182	0.893
ELA	5	4	b4	1.333	2.457	1.133	-1.416	4.483
ELA	5	1	b5				-0.783	-0.783
ELA	5	1	b6				-0.144	-0.144
ELA	5	1	b7				0.931	0.931
ELA	6	285	a1	0.903	0.372	0.866	0.302	2.008
ELA	6	285	b1	-0.845	1.540	-0.947	-6.231	4.177
ELA	6	50	b2	-0.391	1.444	-1.006	-2.634	1.967
ELA	6	11	b3	-0.391	1.444	-1.006	-2.634	1.967
ELA	6	4	b4	-0.848	0.226	-0.769	-1.168	-0.686
ELA	6	2	b5	0.075	0.288	0.075	-0.129	0.279
ELA	6	1	b6				0.937	0.937
ELA	7	269	a1	0.866	0.361	0.816	0.307	2.110
ELA	7	269	b1	-1.001	1.789	-1.002	-8.979	4.827
ELA	7	57	b2	0.655	2.131	0.192	-2.580	5.104
ELA	7	22	b3	0.655	2.131	0.192	-2.580	5.104
ELA	7	4	b4	-0.684	1.232	-0.718	-1.811	0.512
ELA	7	2	b5	-0.969	0.029	-0.969	-0.990	-0.948
ELA	7	1	b6				-0.524	-0.524
ELA	7	1	b7				-0.075	-0.075
ELA	7	1	b8				0.637	0.637
ELA	8	250	a1	0.902	0.349	0.888	0.317	2.131
ELA	8	250	b1	-0.839	1.630	-0.838	-9.064	4.055
ELA	8	37	b2	-0.559	1.842	-0.031	-4.834	1.685
ELA	8	14	b3	-0.559	1.842	-0.031	-4.834	1.685
ELA	8	7	b4	0.055	1.883	-0.021	-1.912	3.479
ELA	8	2	b5	-1.445	0.070	-1.445	-1.494	-1.395
ELA	8	2	b6	-1.066	0.258	-1.066	-1.249	-0.884
ELA	8	2	b7	-0.435	0.432	-0.435	-0.740	-0.129
ELA	8	1	b8				-0.004	-0.004
ELA	10	291	a1	0.897	0.352	0.877	0.315	2.206
ELA	10	291	b1	-0.697	1.667	-0.669	-8.762	6.878
ELA	10	52	b2	-0.112	1.210	-0.379	-2.128	2.254
ELA	10	20	b3	-0.112	1.210	-0.379	-2.128	2.254
ELA	10	5	b4	0.662	2.580	-0.420	-1.194	5.163
ELA	10	3	b5	1.602	0.952	1.912	0.534	2.361
ELA	10	3	b6	2.357	0.613	2.007	2.000	3.064

Continued on next page

Subject	Grade	k	Parm	Mean	SD	Median	Min	Max
Math	3	220	a1	1.067	0.419	1.033	0.309	2.261
Math	3	220	b1	-0.488	1.820	-0.415	-8.101	5.216
Math	3	29	b2	-0.553	2.017	-0.555	-4.331	5.590
Math	3	21	b3	0.771	1.916	0.745	-2.496	4.408
Math	3	7	b4	0.402	1.300	-0.059	-1.214	2.425
Math	3	2	b5	0.102	0.643	0.102	-0.353	0.556
Math	3	2	b6	0.663	0.389	0.663	0.388	0.938
Math	3	1	b7				1.990	1.990
Math	3	1	b8				3.308	3.308
Math	4	238	a1	1.066	0.405	1.033	0.335	2.019
Math	4	238	b1	-0.733	2.041	-0.592	-9.398	6.952
Math	4	1	b10				3.458	3.458
Math	4	61	b2	-1.059	1.802	-1.270	-5.098	5.500
Math	4	50	b3	0.351	1.966	-0.112	-2.671	6.078
Math	4	25	b4	0.395	1.337	0.257	-1.743	3.212
Math	4	12	b5	0.861	1.483	0.248	-1.125	4.293
Math	4	5	b6	0.953	1.275	0.987	-0.494	2.909
Math	4	1	b7				-0.029	-0.029
Math	4	1	b8				0.663	0.663
Math	4	1	b9				1.452	1.452
Math	5	218	a1	1.031	0.380	1.014	0.325	2.096
Math	5	218	b1	-0.320	1.880	-0.202	-6.594	4.633
Math	5	44	b2	-0.674	1.877	-0.904	-4.287	5.126
Math	5	31	b3	0.372	1.574	0.286	-3.314	3.382
Math	5	16	b4	1.153	1.858	0.551	-1.916	4.218
Math	5	2	b5	-0.526	1.175	-0.526	-1.357	0.305
Math	5	2	b6	0.319	1.906	0.319	-1.029	1.666
Math	6	187	a1	0.974	0.458	0.911	0.301	2.361
Math	6	187	b1	-0.392	2.109	-0.047	-7.655	6.515
Math	6	32	b2	-0.558	1.987	-1.071	-3.360	4.598
Math	6	25	b3	1.569	2.754	1.276	-2.109	9.515
Math	6	8	b4	2.064	2.753	1.634	-0.899	6.016
Math	7	198	a1	0.883	0.403	0.821	0.303	2.113
Math	7	198	b1	0.171	2.276	0.346	-8.330	6.816
Math	7	30	b2	-0.414	2.433	-0.483	-6.925	6.314
Math	7	19	b3	0.667	2.072	1.006	-4.325	3.870
Math	7	9	b4	1.688	2.509	0.667	-1.350	5.332
Math	7	4	b5	1.373	0.892	1.042	0.732	2.674
Math	7	1	b6				5.836	5.836
Math	7	1	b7				95.918	95.918
Math	8	194	a1	0.871	0.371	0.846	0.261	2.037
Math	8	194	b1	0.150	2.407	0.406	-8.628	8.416

Continued on next page

Subject	Grade	k	Parm	Mean	SD	Median	Min	Max
Math	8	30	b2	-0.715	1.651	-0.759	-4.108	2.962
Math	8	23	b3	1.233	1.270	1.343	-1.351	3.257
Math	8	14	b4	3.414	2.616	3.111	-0.281	9.600
Math	8	3	b5	2.214	1.501	1.795	0.967	3.880
Math	8	2	b6	4.156	3.312	4.156	1.814	6.497
Math	10	189	a1	0.822	0.412	0.748	0.215	2.175
Math	10	189	b1	0.407	2.403	0.297	-6.311	17.676
Math	10	26	b2	0.908	2.204	0.375	-2.744	7.672
Math	10	17	b3	2.658	2.256	1.630	-0.370	7.520
Math	10	3	b4	4.710	1.702	3.974	3.500	6.656

18
Scaling

Scaling is used to transform a test's raw scores to a new scale that is more useful for test score users. Scaled scores should facilitate proper score interpretations, while minimizing misinterpretations and unwarranted inferences. This is often done by attaching content meaning to the scores.[1]

[1] Peterson, Kolen, and Hoover (1989)

Many state assessments add content meaning by anchoring one or more of a test's performance-level cut scores to known scaled-score values. This approach was used for scaling KAP as KAP scaled scores assign performance-level classifications.

The *Standards for Educational and Psychological Measurement*[2] provides the following guidelines regarding scaling:

[2] APA, AERA, NCME, 2014, p. 102

- Standard 5.1. Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scaled scores, as well as their limitations.
- Standard 5.2. The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.
- Standard 5.3. If there is sound reason to believe that specific misinterpretations of a score scale are likely, test users should be explicitly cautioned.

18.1 Total Test Scaled Scores

Individual student scores on the KAP are reported as scaled scores. However, the student scores are initially estimated as IRT ability estimates. Scaled scores are preferable to IRT ability estimates for reporting purposes. IRT ability values have negative and decimal values. By transforming IRT ability values to scaled scores, all reported values are positive integers which have no decimals are thought to be easier for students and parents to understand.

The chapter on IRT calibration gives more information on the IRT model.

KAP scaled scores were derived using the following two-step process. First, a nonlinear transformation converted KAP raw scores into IRT ability estimates. Next, a *linear transformation* converted the IRT ability estimates into scaled scores. Scaled scores were rounded so that they were all positive integers with three digits. IRT ability estimates and scaled scores will be comparable in the future, after they are linked to the base-year scale.

See the chapter on linking for more information about the linking process.

The linear transformation illustrated in the margin occurs in two-dimensional space where the x-axis shows the θ scale and the y-axis shows scaled scores. A single point on the θ scale was anchored to a single scaled-score value. The slope of the transformation function was also set. Of course, multiple lines, each with different slopes, can go through that single point in the two-dimensional space. The variability of the scaled scores can be controlled through selection of a specific slope value. Further details are provided below, but this is the basic method for scaling the KAP tests.

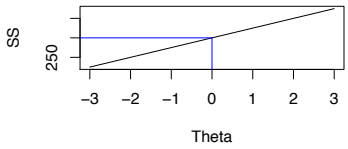


Figure 18.1: Linear Transformation Example

18.1.1 Definition of Scoreability

All students who tested had a scoreable testing event and had an individual student report (ISR) generated. For all items, omits (no responses) were scored as zero.

Rules for excluding students from aggregate reports are discussed in the chapter on score reports.

18.1.2 IRT Ability Estimates

The FlexMIRT[3] computer program was used to derive the IRT ability estimates. FlexMIRT provides a conversion table that maps raw scores to IRT ability estimates. These ability estimates were transformed to scaled scores.

[3] Cai, 2013

Expected *a posteriori* (EAP) ability estimates were used.

18.1.3 Linear Transformation Formulas

KAP scaled scores were obtained through a linear transformation of the IRT ability estimates. The linear transformation anchored one point in two-dimensional space. Specifically, the Level 3 theta-estimate cut score was anchored to the desired Level 3 scaled score cut score. For all KAP subjects and grades, the Level 3 scaled score cut score was anchored to 300 and the slope was set to 25. The underlying formula is $SS_i = A \times \theta_i + C$, where A is the slope (or multiplicative constant) and C is the intercept (or additive constant).

$$A = \frac{SD_{SS}}{SD_\theta}$$

Figure 18.2: Derivation of the Slope Constant: A

$$C = (Cut_{SS} - \frac{SD_{SS}}{SD_\theta} \times Cut_\theta)$$

Figure 18.3: Derivation of the Additive Constant: C

18.1.3.1 θ Cut Scores Coming out of Standard Setting

Table 18.1: Math Theta Cut Scores

Grade	Level 1/2 Cut	Level 2/3 Cut	Level 3/4 Cut
3	-1.225	-0.230	0.906
4	-1.215	0.160	1.375
5	-0.885	0.219	1.245
6	-0.882	0.215	1.340
7	-1.055	0.321	1.980
8	-0.527	0.530	1.968
10	-0.497	0.530	1.830

Table 18.2: ELA Theta Cut Scores

Grade	Level 1/2 Cut	Level 2/3 Cut	Level 3/4 Cut
3	-1.015	-0.050	1.020
4	-1.457	-0.275	1.107
5	-1.085	-0.064	0.952
6	-0.756	0.181	1.594
7	-0.800	0.219	1.610
8	-0.940	0.495	1.850
10	-0.785	0.465	1.800

18.1.3.2 Slope and Intercept Constants

Table 18.3: Slopes and Intercepts for Deriving Math Scaled Scores

Grade	A	C
3	25	305.7500
4	25	296.0000
5	25	294.5312
6	25	294.6250
7	25	291.9750
8	25	286.7500
10	25	286.7500

18.1.3.3 Scaled-Score Cut Scores

18.1.4 Lowest and Highest Obtainable Scaled Scores

All KAP mathematics and ELA tests have a lowest obtainable scaled score (LOSS) of 220 and a highest obtainable scaled score (HOSS) of 380.

Grade	A	C
3	25	301.2500
4	25	306.8675
5	25	301.5875
6	25	295.4750
7	25	294.5250
8	25	287.6250
10	25	288.3750

Table 18.4: Slopes and Intercepts for Deriving ELA Scaled Scores

Grade	Level 1/2 Cut	Level 2/3 Cut	Level 3/4 Cut
3	276	300	329
4	266	300	331
5	273	300	326
6	273	300	329
7	266	300	342
8	274	300	336
10	275	300	333

Table 18.5: Math SS Cut Scores

18.1.5 Rounding

The linearly transformed scaled scores were rounded to the nearest integer value for reporting purposes. Values greater than or equal to 0.50 were rounded up. Values less than 0.50 were rounded down.

18.1.6 Decimal Raw Scores

Because some KAP items allowed decimal scores, total raw scores could end in decimals too. FlexMIRT's EAP ability estimates provided conversions for all possible decimal raw scores. However, the conversion tables had some odd properties. First, many conversions were produced; for example, a sixty-point test had over 1,000 raw-score to theta-score conversions. Second, FlexMIRT only reported raw scores to two decimal places, whereas actual raw scores could differ beyond the third decimal place. Because of this, some decimal raw scores seemed to appear multiple times in the conversion tables. Finally, the conversions were not monotonically increasing; in other words, as raw scores increased, scaled scores did not always increase.

For reporting purposes, decimal raw scores were rounded to the nearest whole number; raw scores ending in exactly 0.5 were rounded up. The scaled score reported for each integer raw score was the weighted average of all scaled scores occurring between each raw score's *real limits*. The linearly transformed scaled scores were rounded to the nearest integer value for reporting purposes.

For any raw score x, the real limits are $x-0.5$ and $x+0.5$. For example, if $x = 7$, the real limits are 6.5 and 7.5

Table 18.6: ELA SS Cut Scores

Grade	Level 1/2 Cut	Level 2/3 Cut	Level 3/4 Cut
3	276	300	329
4	266	300	331
5	273	300	326
6	273	300	329
7	266	300	342
8	274	300	336
10	275	300	333

18.1.7 Example Raw-Score to Scaled-Score Table

A sample raw-score (RS) to scaled-score (SS) table is provided below. Because of their size, complete raw-score to scaled-score tables are documented in an appendix.

Table 18.7: Sample RS-SS Table

Year	Subject	Grade	FormID	RS	SS	SEM
2015	M	3	4192	0	220	13.3
2015	M	3	4192	1	221	12.8
2015	M	3	4192	2	226	12.3
2015	M	3	4192	3	231	11.7
2015	M	3	4192	4	236	11.2
2015	M	3	4192	5	240	10.8
2015	M	3	4192	6	244	10.3
2015	M	3	4192	7	248	9.9
2015	M	3	4192	8	251	9.5
2015	M	3	4192	9	255	9.2

18.2 Claim Scores

With a few minor exceptions, the same process was used to derive scaled scores for math and ELA claim scores. Using FlexMIRT, the item parameters derived from the scaling of the total test scores were used to derive ability estimates for each claim. As with the total scores, IRT ability estimates were based on EAP. Linear transformations were then applied to each claim scaled score. Instead of anchoring the Level 3 θ cut score to a particular scaled score, the population latent ability mean in the scaling of the total test scores was anchored to a scaled-score value of 110 and the slope of the transformation was set to 2. LOSS and HOSS values were set to 106 and 114, respectively, for all claim scores. Scaled scores were rounded to the nearest whole number (decimals of 0.5 were rounded up). As described above, decimal raw scores for claims were rounded to the nearest integer, and weighted averages of the scaled scores within the real limits of the raw

In FlexMIRT, the mean of θ in the latent population is 0.

scores were reported.

19
Linking

This chapter describes the test design elements and associated data collection and analysis procedures that relate to linking scores from different KAP test forms.

In large-scale assessment programs, different item sets can be used on test forms both within years and across years. Linking the scores from these different test forms puts the form scores on a common scale of measurement and ensures that all forms for a given grade-level and subject-area tests provide comparable scores. This means that students will not have an unfair advantage or disadvantage simply because the test form they took was easier or harder than the test forms taken by other students.

There was no need for across year linking this year because this was the first administration of KAP

19.1 Test Design Elements

19.1.1 Versions of the Assessment

There are multiple versions of the KAP assessment. By far, the dominant administration format for KAP was via the KITE computer testing platform which can be used on PCs with Windows, Macs, Chromebooks, and iPads. An extremely small number of students took KAP using an alternate testing format (Braille, large-print, paper and pencil). No grade-level or subject area test had more than 45 students taking an alternate test format. Although the paper and pencil format was the most common alternative, no more than 25 students took paper and pencil format at any given grade.

Sample sizes were too small to undertake an comparability studies for the alternative testing formats

19.1.2 Alternate Forms

KAP tests use multiple test forms. In 2015 eight forms were used for all ELA and math grade level tests. Each test form was con-

structed to include a total of 70 items. This number was selected, in part, because the future multistage adaptive testing (MST) format planned for KAP will have four stages with 25, 15, 15, and 15 items in each stage, respectively, for a total of 70 items.

Although the KAP tests included 70 items, fewer items contributed to student test scores. Items that contributed to student scores were selected on a form-by-form basis once their item statistics were reviewed after testing.

Test developers had more confidence that some items would perform better than others. These items were positioned at the start of the test (i.e., the first 25 items) and appeared on all test forms. In the jargon of test linking, these were *common* items across forms. The remaining 45 items were generally unique across the eight test forms, although a few items were occasionally repeated on some form pairs. The following figure demonstrates the basic structure of the tests.

Operational items could not be designated in advance of testing as is typical because a cyber-attack during the prior administration year. Due to this attack, the psychometric properties of many of the test items were unknown when tests were constructed. Consequently, operational items were selected after the testing window ended and item statistics were available.

Table 19.1: KAP Test Design

Form/Items	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9
A	*	*							
B	*		*						
C	*			*					
D	*				*				
E	*					*			
F	*						*		
G	*							*	
H	*								*

19.1.3 Test Length

As noted above, some items did not contribute to student scores. Most often this was because the items had marginal performance statistics (e.g., poorly discriminating items) or item statistics suggested that item modifications were required (e.g., there were multiple correct responses). Items could be excluded for other reasons as well. Some off-grade testing was done in the interest of expanding the difficulty of the item pools to prepare for the transition to the MST administration format. Such off-grade items were excluded this year. A few items that did not display properly on computer screens during the testing were also excluded.

Off-grade items might be used in the future but should measure an appropriate on-grade standard.

Because items were excluded on a form-by-form basis, some forms had more total points possible than other forms. A decision was made to report scores from each form on as many items as possible. This way, no additional items had to be excluded in order to equalize the

points possible across forms. However, this suggests that the form scores do not meet the strict criteria needed for the strongest form of linking, known as equating. For equating, it should be a matter of indifference to the examinee which form of the assessment he or she takes. Because longer tests tend to have smaller standard errors of measurement, a student near the KAP Level 3 cut score might actually prefer taking a test form with more items. Hence, form scores would not be completely exchangeable from the students' perspective.

The minimum and maximum point ranges across forms are summarized in the following tables. Reliability indices and conditional standard errors of measurement (CSEMs) for each test form are documented in the reliability chapter. While some tests were more similar in length and reliability (e.g., Math Grade 10, ELA Grade 10), others had considerably more variation across forms (e.g., Math Grade 4 and ELA Grade 7).

The total points possible for all test forms are given in the tables at the end of this section.

Subject	Grade	Min. Pts.	Max. Pts.
Math	3	51	66
Math	4	46	68
Math	5	51	68
Math	6	49	69
Math	7	47	65
Math	8	47	63
Math	10	52	60

Table 19.2: Minimum and Maximum Points across Forms by Grade for Math

Subject	Grade	Min. Pts.	Max. Pts.
ELA	3	58	77
ELA	4	58	80
ELA	5	61	77
ELA	6	57	73
ELA	7	58	82
ELA	8	55	76
ELA	10	71	85

Table 19.3: Minimum and Maximum Points across Forms by Grade for ELA

19.1.4 Content Distribution

Another factor that affects the strength of any score linkage is the similarity of content on the test forms. Minimum and maximum content percentages are summarized in the following tables. In Math, the initial build targets were to have 65% to 75% of the items in Claim 1 and the remaining 25% to 35% of items in Claims 2, 3 and 4 (roughly 8% to 12% in each, respectively). Final operational Claim 1 items selected for math resulted in a low to high range of 66% to 76%. The initial target for ELA was to have 60% to 65% of items in Claim 1 and the remaining 30% to 35% of items in Claim 2. Final operational Claim 1 items selected for ELA resulted in a low to high range of 56% to 71%. The ELA range is a little wider, which is not unusual as the testlet structures used on this assessment make it more challenging to hit content targets precisely.

The claim proportions shown below are relative to the total points possible. Claim proportions for all tests forms are given in the tables at the end of this section.

Subject	Grade	Min. Prop.	Max. Prop.
Math	3	0.68	0.72
Math	4	0.67	0.74
Math	5	0.66	0.73
Math	6	0.67	0.76
Math	7	0.67	0.74
Math	8	0.71	0.75
Math	10	0.67	0.75

Table 19.4: Minimum and Maximum Claim 1 Proportions across Form by Grade for Math

Subject	Grade	Min. Prop.	Max. Prop.
Math	3	0.10	0.13
Math	4	0.07	0.12
Math	5	0.08	0.12
Math	6	0.08	0.12
Math	7	0.11	0.13
Math	8	0.06	0.10
Math	10	0.08	0.15

Table 19.5: Minimum and Maximum Claim 2 Proportions across Form by Grade for Math

Subject	Grade	Min. Prop.	Max. Prop.
Math	3	0.06	0.12
Math	4	0.08	0.14
Math	5	0.10	0.12
Math	6	0.09	0.12
Math	7	0.09	0.12
Math	8	0.08	0.11
Math	10	0.07	0.11

Table 19.6: Minimum and Maximum Claim 3 Proportions across Form by Grade for Math

Subject	Grade	Min. Prop.	Max. Prop.
Math	3	0.08	0.12
Math	4	0.07	0.12
Math	5	0.09	0.12
Math	6	0.06	0.12
Math	7	0.04	0.10
Math	8	0.06	0.11
Math	10	0.05	0.12

Table 19.7: Minimum and Maximum Claim 4 Proportions across Form by Grade for Math

Subject	Grade	Min. Prop.	Max. Prop.
Math	3	0.28	0.32
Math	4	0.26	0.33
Math	5	0.27	0.34
Math	6	0.24	0.33
Math	7	0.26	0.33
Math	8	0.25	0.29
Math	10	0.25	0.33

Table 19.8: Minimum and Maximum Claim 2, 3 and 4 Proportions across Form by Grade for Math

Subject	Grade	Min. Prop.	Max. Prop.
ELA	3	0.61	0.67
ELA	4	0.62	0.71
ELA	5	0.60	0.66
ELA	6	0.59	0.68
ELA	7	0.59	0.68
ELA	8	0.59	0.69
ELA	10	0.56	0.71

Table 19.9: Minimum and Maximum Claim 1 Proportions across Form by Grade for ELA

Subject	Grade	Min. Prop.	Max. Prop.
ELA	3	0.24	0.36
ELA	4	0.25	0.39
ELA	5	0.20	0.37
ELA	6	0.23	0.41
ELA	7	0.21	0.40
ELA	8	0.19	0.37
ELA	10	0.23	0.42

Table 19.10: Minimum and Maximum Claim 1 Literary Proportions across Form by Grade for ELA

Subject	Grade	Min. Prop.	Max. Prop.
ELA	3	0.26	0.41
ELA	4	0.25	0.42
ELA	5	0.26	0.43
ELA	6	0.25	0.42
ELA	7	0.22	0.39
ELA	8	0.23	0.42
ELA	10	0.24	0.39

Table 19.11: Minimum and Maximum Claim 1 Informational Proportions across Form by Grade for ELA

Subject	Grade	Min. Prop.	Max. Prop.
ELA	3	0.33	0.39
ELA	4	0.29	0.38
ELA	5	0.34	0.40
ELA	6	0.32	0.41
ELA	7	0.32	0.41
ELA	8	0.31	0.41
ELA	10	0.29	0.44

Table 19.12: Minimum and Maximum Claim 2 Proportions across Form by Grade for ELA

The previous tables only summarized the minimum and maximum test form lengths and content distribution proportions. Complete results for these test features for all forms follow.

Table 19.13: Content Distribution for Grade 3 Math Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_2	Claim_3	Claim_4	Claim_234
Math	3	A	64	0.69	0.12	0.09	0.09	0.31
Math	3	B	51	0.71	0.10	0.12	0.08	0.29
Math	3	C	63	0.70	0.13	0.06	0.11	0.30
Math	3	D	66	0.68	0.11	0.11	0.11	0.32
Math	3	E	52	0.69	0.12	0.08	0.12	0.31
Math	3	F	61	0.72	0.10	0.08	0.10	0.28
Math	3	G	64	0.70	0.11	0.09	0.09	0.30
Math	3	H	63	0.70	0.13	0.08	0.10	0.30

k = Points Possible

Table 19.14: Content Distribution for Grade 4 Math Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_2	Claim_3	Claim_4	Claim_234
Math	4	A	67	0.67	0.10	0.10	0.12	0.33
Math	4	B	46	0.74	0.07	0.09	0.11	0.26
Math	4	C	67	0.69	0.10	0.10	0.10	0.31
Math	4	D	66	0.68	0.11	0.14	0.08	0.32
Math	4	E	51	0.71	0.12	0.08	0.10	0.29
Math	4	F	68	0.68	0.10	0.10	0.12	0.32
Math	4	G	67	0.69	0.10	0.13	0.07	0.31
Math	4	H	64	0.69	0.09	0.11	0.11	0.31

k = Points Possible

Table 19.15: Content Distribution for Grade 5 Math Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_2	Claim_3	Claim_4	Claim_234
Math	5	A	68	0.66	0.10	0.12	0.12	0.34
Math	5	B	52	0.69	0.10	0.10	0.12	0.31
Math	5	C	65	0.68	0.09	0.12	0.11	0.32
Math	5	D	65	0.69	0.09	0.11	0.11	0.31
Math	5	E	51	0.73	0.08	0.10	0.10	0.27
Math	5	F	62	0.71	0.08	0.10	0.11	0.29
Math	5	G	64	0.69	0.11	0.11	0.09	0.31
Math	5	H	65	0.68	0.12	0.11	0.09	0.32

k = Points Possible

Table 19.16: Content Distribution for Grade 6 Math Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_2	Claim_3	Claim_4	Claim_234
Math	6	A	69	0.67	0.10	0.12	0.12	0.33
Math	6	B	50	0.68	0.12	0.10	0.10	0.32
Math	6	C	64	0.67	0.11	0.11	0.11	0.33
Math	6	D	64	0.72	0.09	0.09	0.09	0.28
Math	6	E	49	0.76	0.08	0.10	0.06	0.24
Math	6	F	58	0.71	0.10	0.10	0.09	0.29
Math	6	G	65	0.68	0.09	0.12	0.11	0.32
Math	6	H	53	0.70	0.09	0.11	0.09	0.30

k = Points Possible

Table 19.17: Content Distribution for Grade 7 Math Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_2	Claim_3	Claim_4	Claim_234
Math	7	A	60	0.70	0.12	0.12	0.07	0.30
Math	7	B	47	0.74	0.11	0.11	0.04	0.26
Math	7	C	63	0.70	0.13	0.11	0.06	0.30
Math	7	D	62	0.71	0.13	0.10	0.06	0.29
Math	7	E	51	0.69	0.12	0.10	0.10	0.31
Math	7	F	65	0.69	0.12	0.09	0.09	0.31
Math	7	G	61	0.69	0.11	0.10	0.10	0.31
Math	7	H	48	0.67	0.12	0.10	0.10	0.33

k = Points Possible

Table 19.18: Content Distribution for Grade 8 Math Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_2	Claim_3	Claim_4	Claim_234
Math	8	A	63	0.73	0.10	0.08	0.10	0.27
Math	8	B	49	0.71	0.10	0.08	0.10	0.29
Math	8	C	56	0.71	0.07	0.11	0.11	0.29
Math	8	D	61	0.75	0.07	0.08	0.10	0.25
Math	8	E	48	0.75	0.08	0.10	0.06	0.25
Math	8	F	59	0.75	0.10	0.08	0.07	0.25
Math	8	G	61	0.72	0.07	0.11	0.10	0.28
Math	8	H	47	0.74	0.06	0.11	0.09	0.26

k = Points Possible

Table 19.19: Content Distribution for Grade 10 Math Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_2	Claim_3	Claim_4	Claim_234
Math	10	A	57	0.74	0.09	0.11	0.07	0.26
Math	10	B	55	0.67	0.15	0.09	0.09	0.33
Math	10	C	59	0.69	0.14	0.07	0.10	0.31
Math	10	D	58	0.74	0.10	0.10	0.05	0.26
Math	10	E	57	0.74	0.09	0.07	0.11	0.26
Math	10	F	56	0.71	0.11	0.09	0.09	0.29
Math	10	G	60	0.75	0.08	0.08	0.08	0.25
Math	10	H	52	0.67	0.13	0.08	0.12	0.33

k = Points Possible

Table 19.20: Content Distribution for Grade 3 ELA Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_1_Info	Claim_1_Lit	Claim_2
ELA	3	A	74	0.61	0.28	0.32	0.39
ELA	3	B	58	0.67	0.36	0.31	0.33
ELA	3	C	73	0.62	0.27	0.34	0.38
ELA	3	D	58	0.66	0.36	0.29	0.34
ELA	3	E	77	0.65	0.29	0.36	0.35
ELA	3	F	76	0.61	0.26	0.34	0.39
ELA	3	G	76	0.64	0.41	0.24	0.36
ELA	3	H	74	0.65	0.28	0.36	0.35

k = Points Possible

Table 19.21: Content Distribution for Grade 4 ELA Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_1_Info	Claim_1_Lit	Claim_2
ELA	4	A	80	0.64	0.25	0.39	0.36
ELA	4	B	58	0.71	0.36	0.34	0.29
ELA	4	C	71	0.69	0.41	0.28	0.31
ELA	4	D	60	0.67	0.33	0.33	0.33
ELA	4	E	74	0.62	0.36	0.26	0.38
ELA	4	F	77	0.64	0.26	0.38	0.36
ELA	4	G	72	0.67	0.42	0.25	0.33
ELA	4	H	69	0.68	0.29	0.39	0.32

k = Points Possible

Table 19.22: Content Distribution for Grade 5 ELA Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_1_Info	Claim_1_Lit	Claim_2
ELA	5	A	76	0.63	0.26	0.37	0.37
ELA	5	B	61	0.66	0.34	0.31	0.34
ELA	5	C	74	0.64	0.43	0.20	0.36
ELA	5	D	61	0.64	0.33	0.31	0.36
ELA	5	E	77	0.62	0.38	0.25	0.38
ELA	5	F	71	0.65	0.41	0.24	0.35
ELA	5	G	77	0.60	0.26	0.34	0.40
ELA	5	H	75	0.64	0.39	0.25	0.36

k = Points Possible

Table 19.23: Content Distribution for Grade 6 ELA Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_1_Info	Claim_1_Lit	Claim_2
ELA	6	A	73	0.64	0.27	0.37	0.36
ELA	6	B	57	0.63	0.33	0.30	0.37
ELA	6	C	69	0.65	0.42	0.23	0.35
ELA	6	D	58	0.64	0.36	0.28	0.36
ELA	6	E	73	0.59	0.25	0.34	0.41
ELA	6	F	72	0.64	0.29	0.35	0.36
ELA	6	G	69	0.68	0.28	0.41	0.32
ELA	6	H	66	0.64	0.39	0.24	0.36

k = Points Possible

Table 19.24: Content Distribution for Grade 7 ELA Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_1_Info	Claim_1_Lit	Claim_2
ELA	7	A	72	0.68	0.28	0.40	0.32
ELA	7	B	58	0.67	0.38	0.29	0.33
ELA	7	C	74	0.66	0.39	0.27	0.34
ELA	7	D	58	0.60	0.36	0.24	0.40
ELA	7	E	80	0.61	0.24	0.38	0.39
ELA	7	F	82	0.59	0.35	0.23	0.41
ELA	7	G	80	0.59	0.22	0.36	0.41
ELA	7	H	76	0.59	0.38	0.21	0.41

k = Points Possible

Table 19.25: Content Distribution for Grade 8 ELA Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_1_Info	Claim_1_Lit	Claim_2
ELA	8	A	75	0.63	0.25	0.37	0.37
ELA	8	B	57	0.65	0.32	0.33	0.35
ELA	8	C	68	0.63	0.28	0.35	0.37
ELA	8	D	55	0.69	0.38	0.31	0.31
ELA	8	E	64	0.61	0.42	0.19	0.39
ELA	8	F	76	0.63	0.28	0.36	0.37
ELA	8	G	73	0.60	0.23	0.37	0.40
ELA	8	H	70	0.59	0.26	0.33	0.41

k = Points Possible

Table 19.26: Content Distribution for Grade 10 ELA Claim Scores

Subject	Grade	Form	k	Claim_1	Claim_1_Info	Claim_1_Lit	Claim_2
ELA	10	A	71	0.62	0.27	0.35	0.38
ELA	10	B	74	0.69	0.39	0.30	0.31
ELA	10	C	72	0.71	0.29	0.42	0.29
ELA	10	D	85	0.60	0.24	0.36	0.40
ELA	10	E	71	0.59	0.25	0.34	0.41
ELA	10	F	82	0.56	0.33	0.23	0.44
ELA	10	G	83	0.58	0.33	0.25	0.42
ELA	10	H	83	0.60	0.25	0.35	0.40

k = Points Possible

19.2 Linking Procedure

19.2.1 Data Collection Design

As noted above, all KAP test forms included a set of common items. This design is most often used when groups taking each form are known to be nonequivalent in their ability. Additionally, all eight test forms were randomly assigned to general education students by the KITE testing engine. One test form was used to test all students who had special needs. One might expect that the group of students who took this form would be somewhat lower in ability, thus making the common items appear somewhat harder on this form. The special needs students were not included in the calibration samples. This should have resulted in *randomly equivalent groups* of students in the calibration samples. This, in conjunction with the aforementioned common items, should have resulted in a robust calibration design.

19.2.2 Linking Method

Item response theory (IRT) scale linking methodologies can place item parameters and student ability estimates on the same scale when different forms are used. KAP items were scaled using a *concurrent calibration* approach. This approach can be successful when common items are present, when all groups are randomly equivalent, or both.

The success of IRT depends on how well the IRT assumptions are met. See the calibration chapter for an evaluation of several key IRT assumptions.

The stability of the common items are illustrated in the following scatter plots. In these graphs, different marker shapes are used for dichotomous items and polytomous items, denoted as *pv* (short for *p*-value) and *pp* (short for pseudo-*p*-value), respectively, in the graph legend.

For both types of items, the proportion-correct difficulties were normalized and scaled to have a mean of 100 and a standard deviation of 10. Note that harder items (i.e., those with lower mean scores) have higher values on this scale.

For example, a proportion-correct mean of 0.8413 would have a normalized scale value of 90 while a proportion-correct mean of 0.1587 would have a value of 110.

If each pair of forms had their item difficulties plotted separately, 28 scatter plots (8 choose 2 = 28, per test) would have been required. To save space, all possible combinations are plotted on a single graph and different color markers used to designate one of the forms. Designating both forms would have required 28 colors and would have likely made the graph too difficult to interpret. Fortunately, all of the items appeared very stable rendering specific form pair designations moot.

The dominant feature of these graphs is that the Form A difficulties were generally shifted to the right of the identity line indicating they were more difficult, which was expected. The shift to the right was not always a uniform distance relative to the identity line. Its possible that some item modifications used for special needs students interacted with item difficulty in these cases. Also, the Form A shift was not as dramatic at Grade 10. Finally, in math Grade 10, some Form B

Although not included during item calibration, students with special needs are included in the following scatter plots. If these plots had been produced without the special needs students, it seems highly likely that the Form A markers would have been near the markers for the other forms.

difficulties were also shifted slightly to the right.

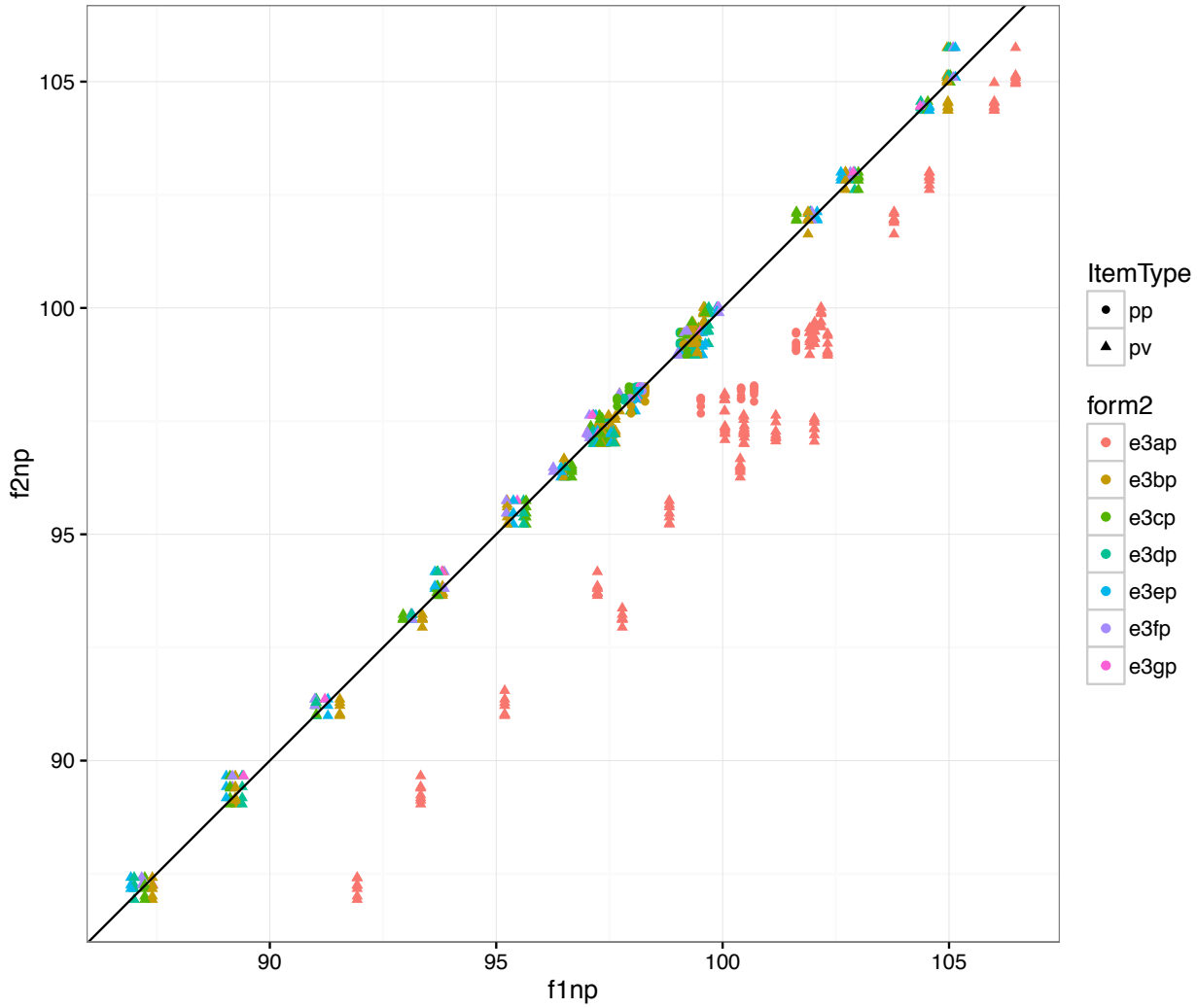


Figure 19.1: ELA Grade 3 Item Stability Plot

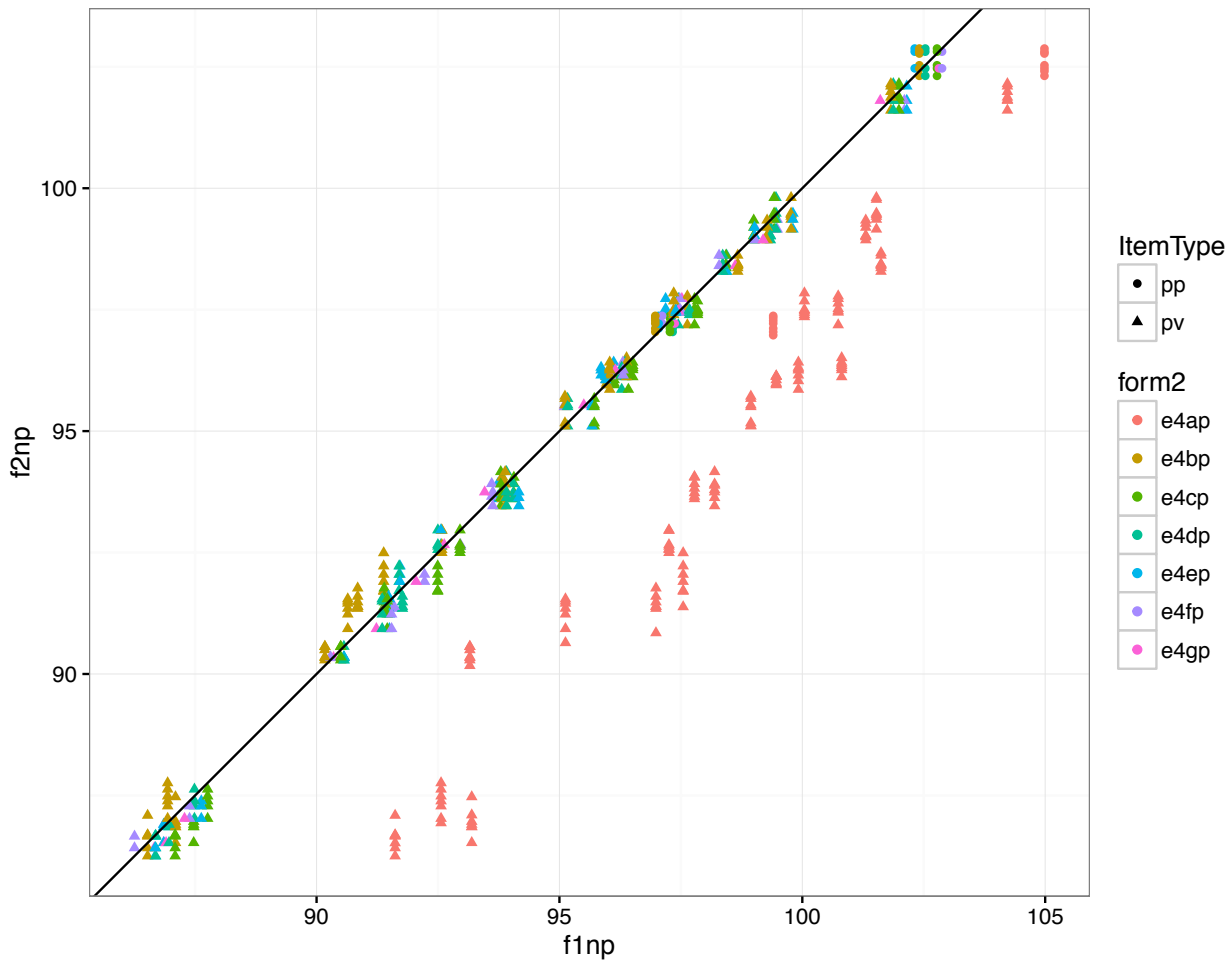


Figure 19.2: ELA Grade 4 Item Stability Plot

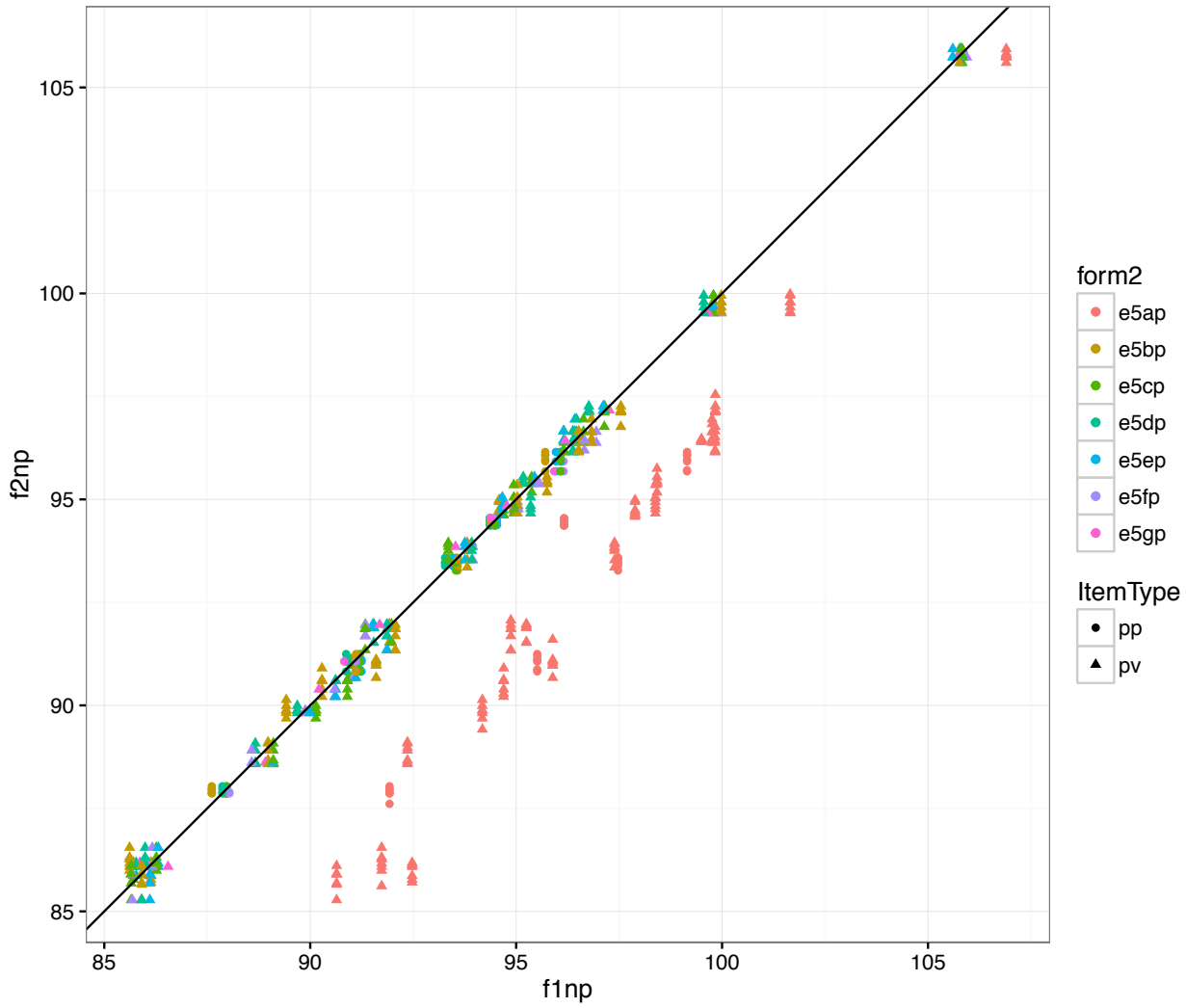


Figure 19.3: ELA Grade 5 Item Stability Plot

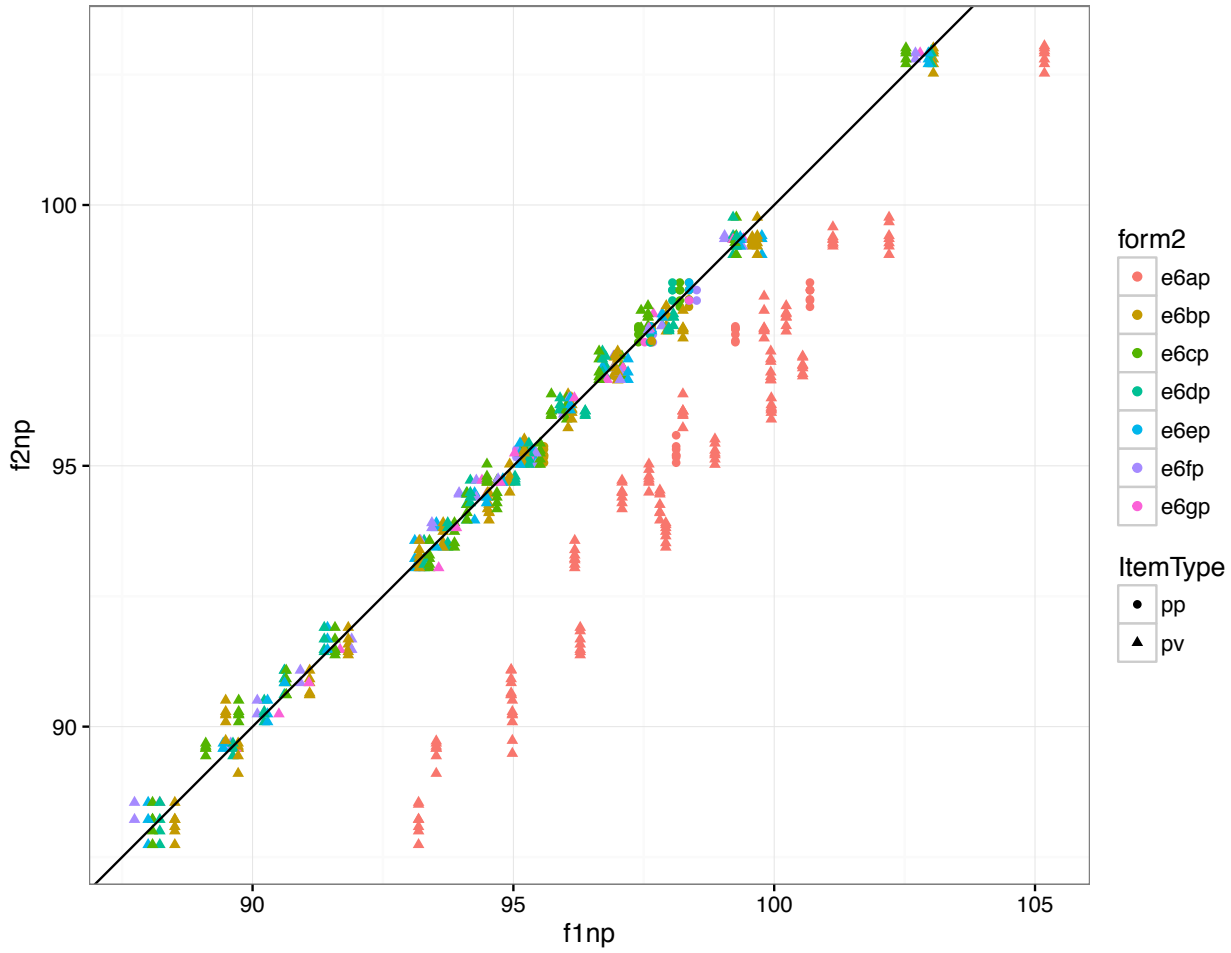


Figure 19.4: ELA Grade 6 Item Stability Plot

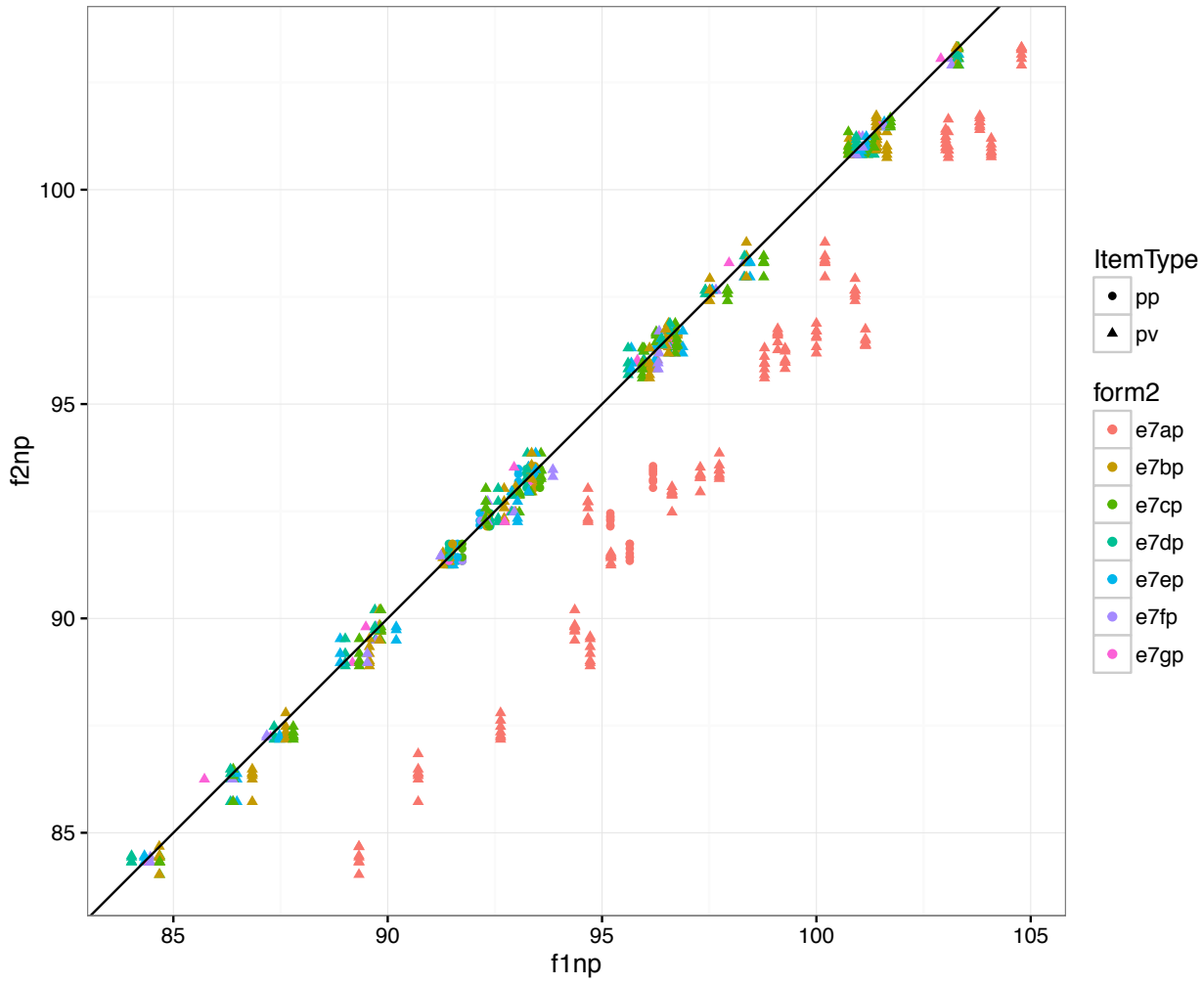


Figure 19.5: ELA Grade 7 Item Stability Plot

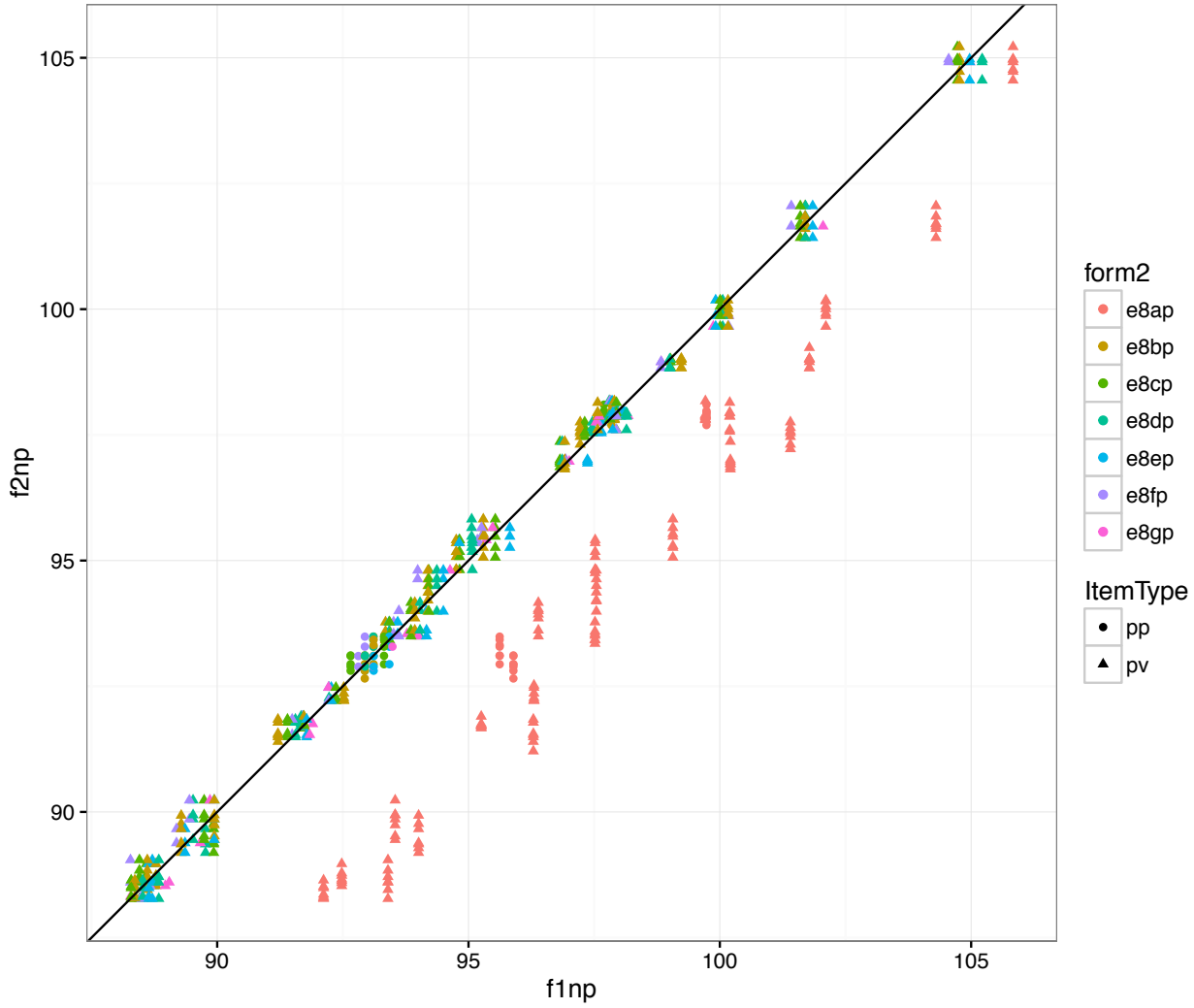


Figure 19.6: ELA Grade 8 Item Stability Plot

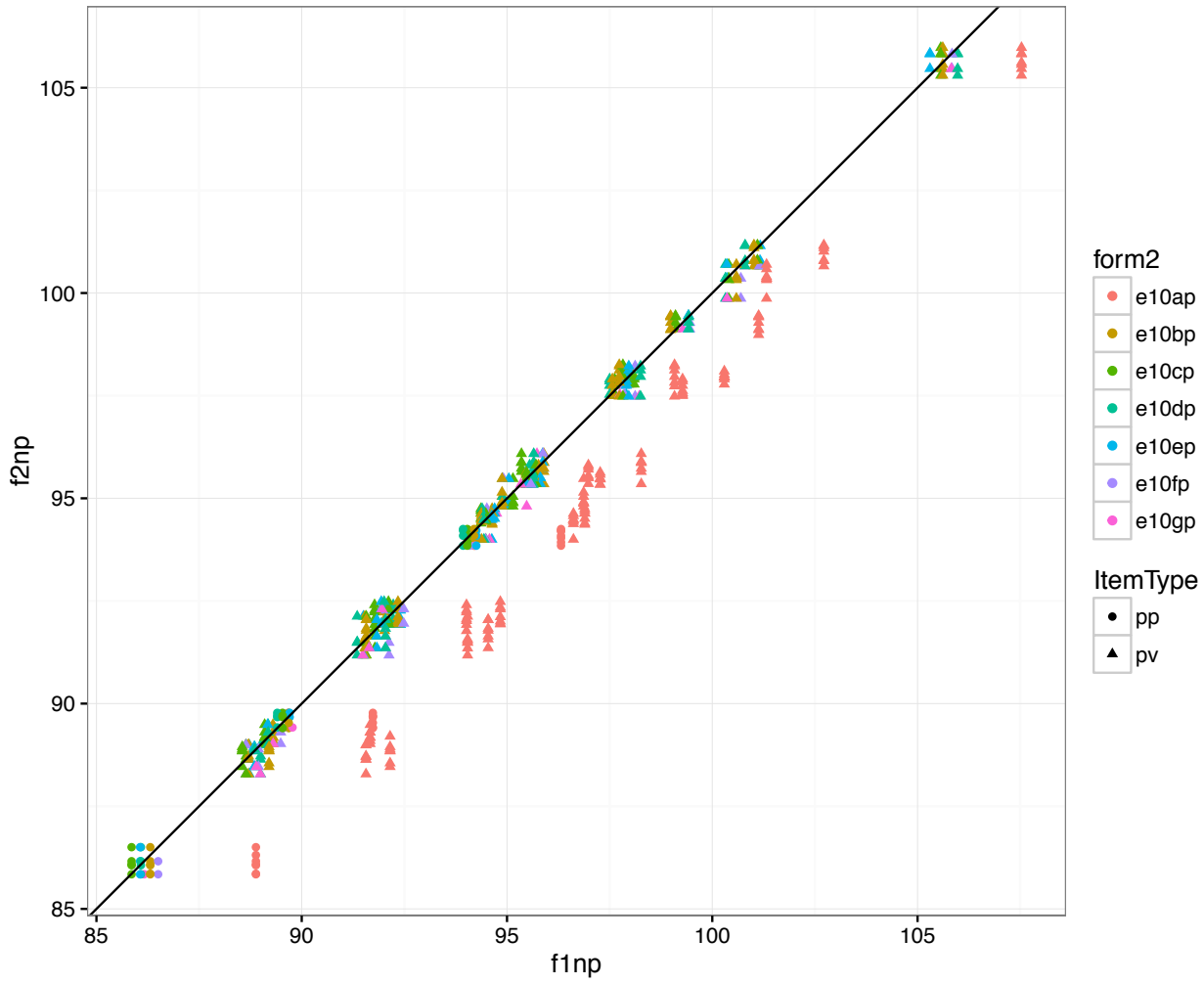


Figure 19.7: ELA Grade 10 Item Stability Plot

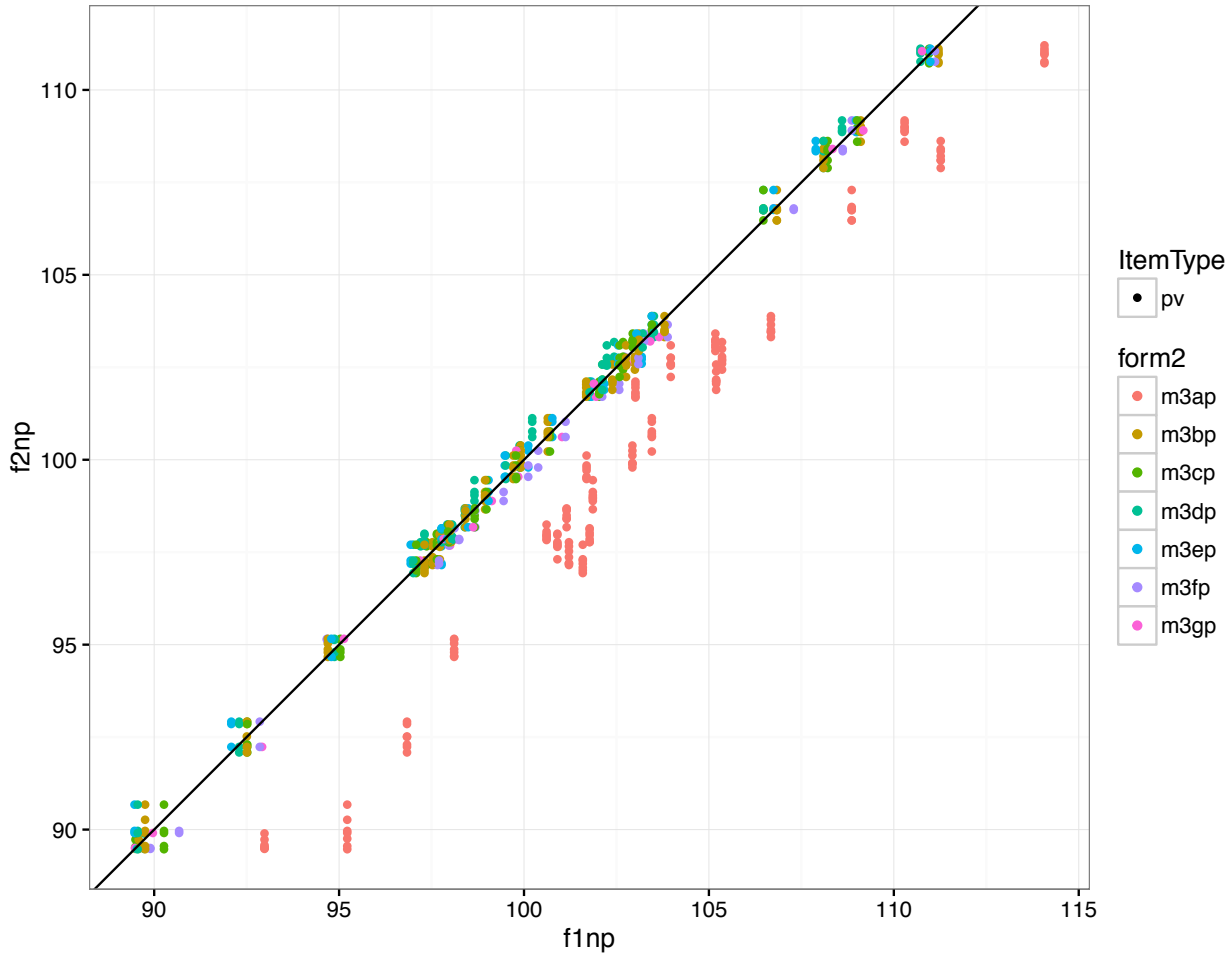


Figure 19.8: Math Grade 3 Item Stability Plot

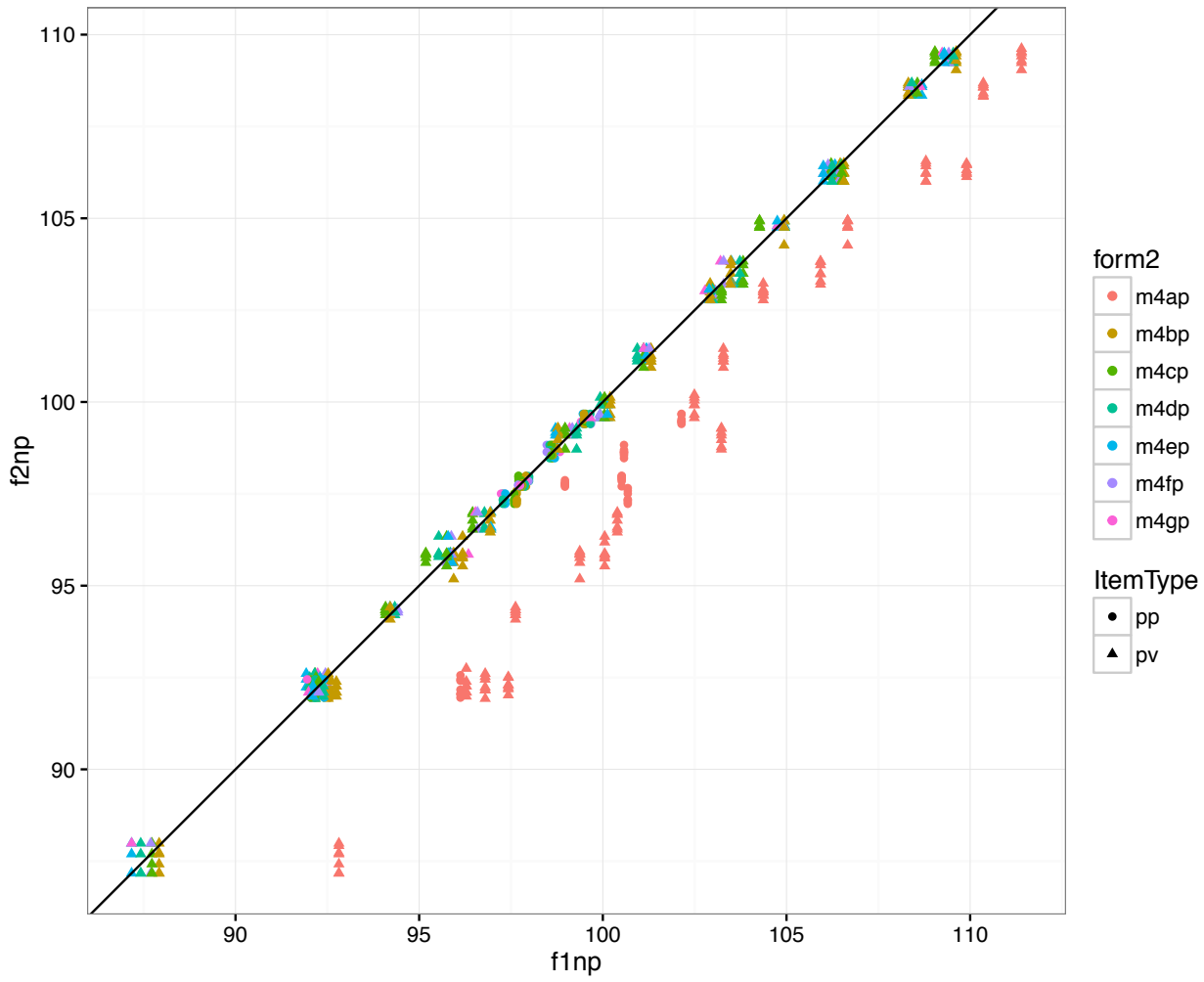


Figure 19.9: Math Grade 4 Item Stability Plot

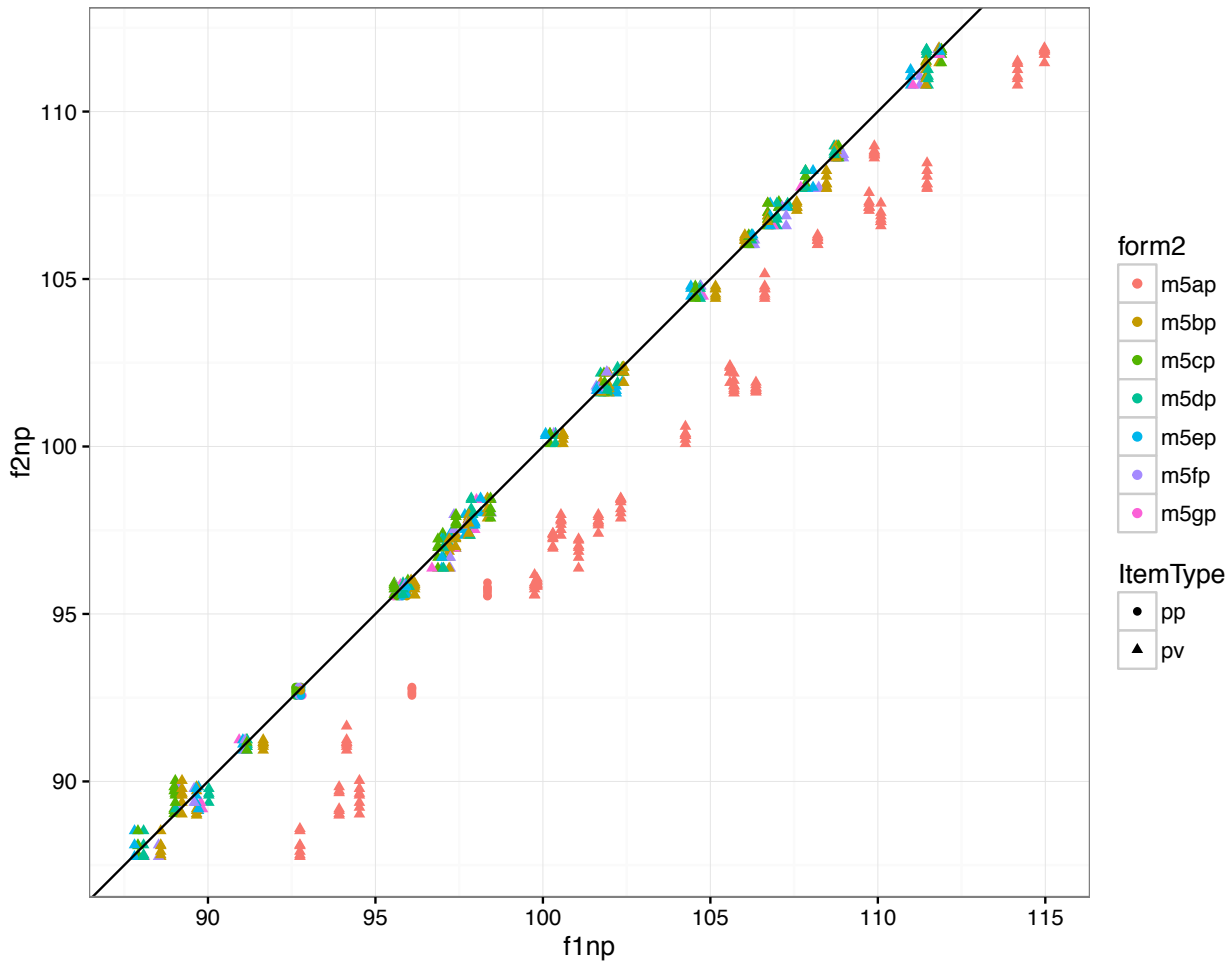


Figure 19.10: Math Grade 5 Item Stability Plot

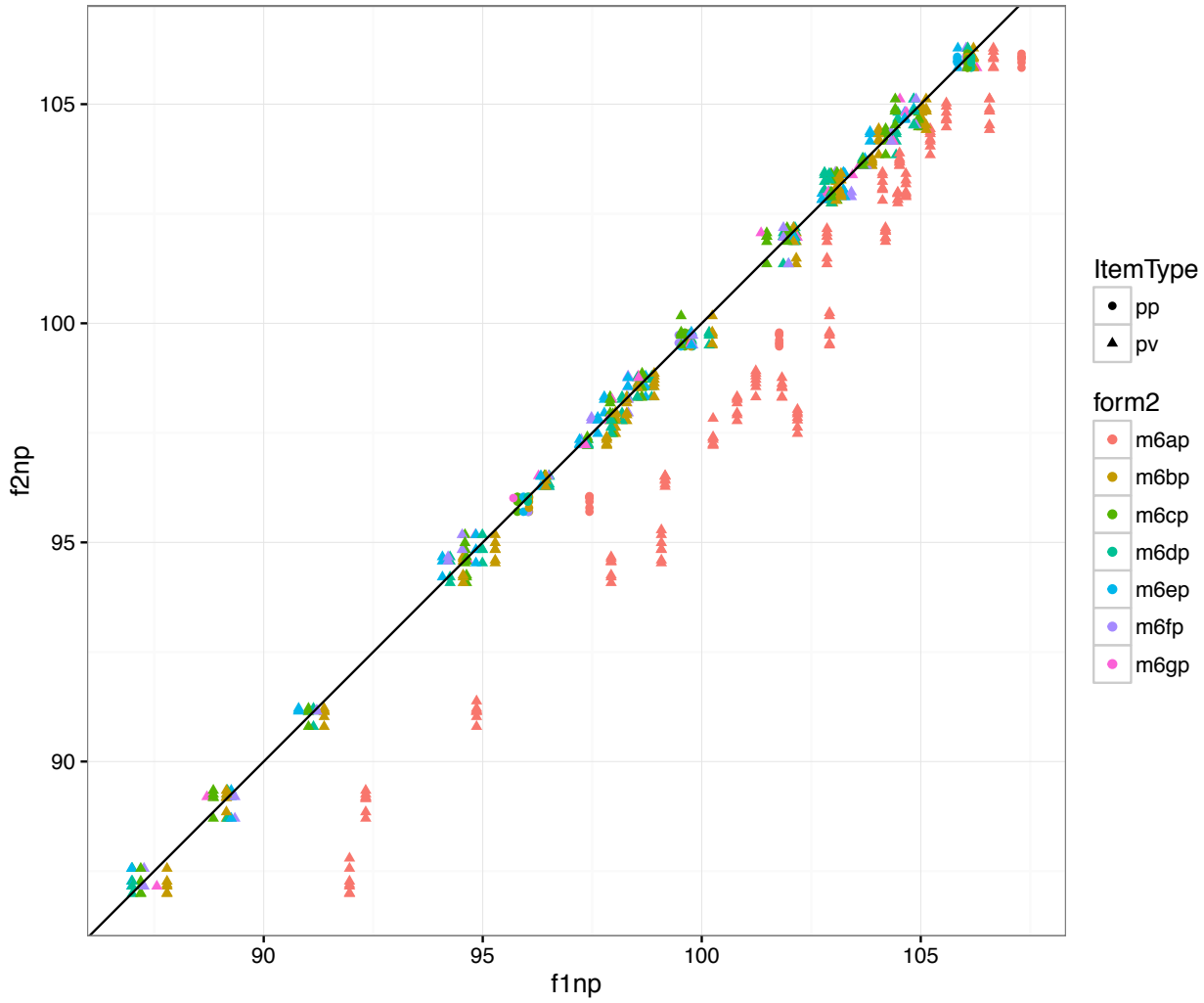


Figure 19.11: Math Grade 6 Item Stability Plot

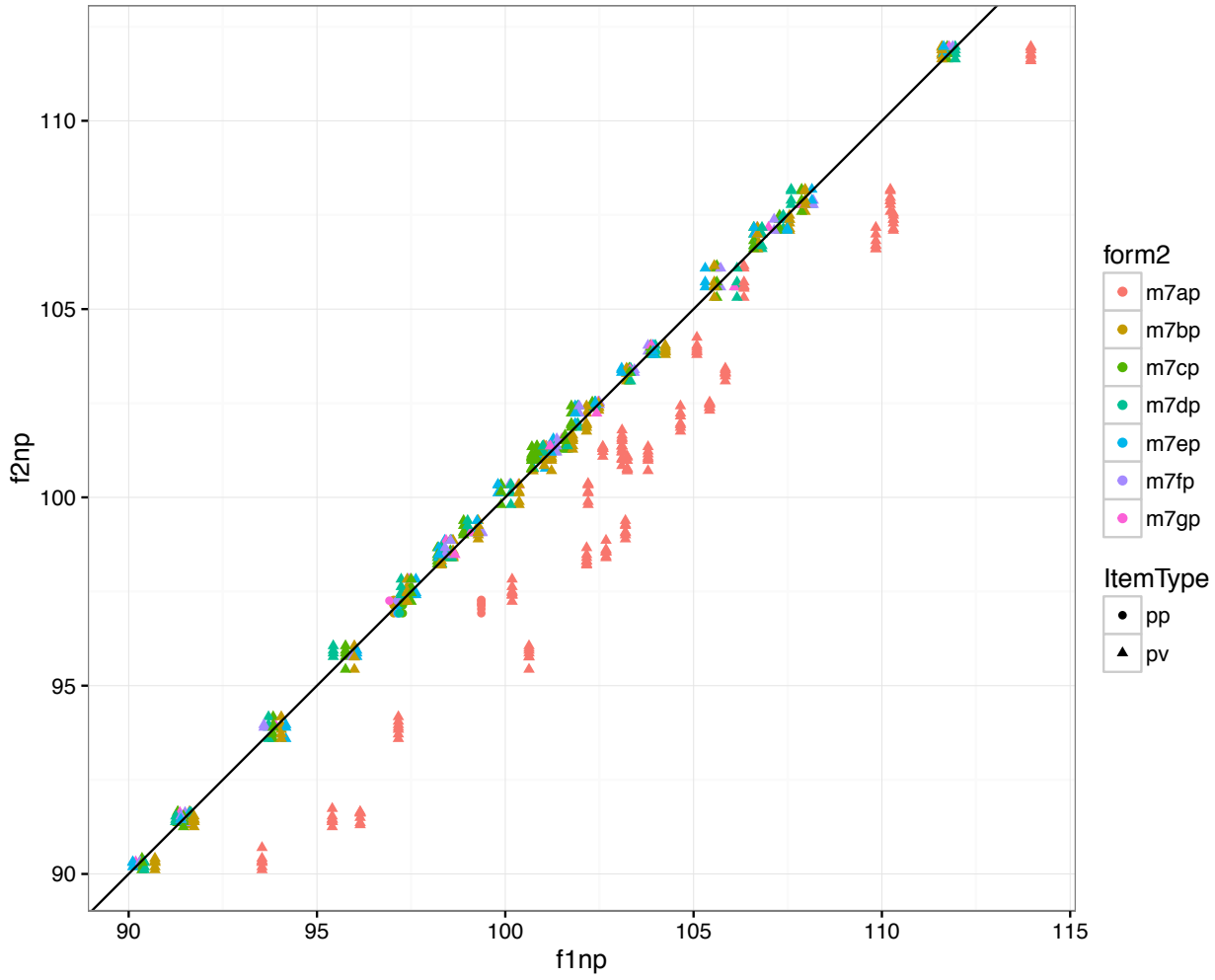


Figure 19.12: Math Grade 7 Item Stability Plot

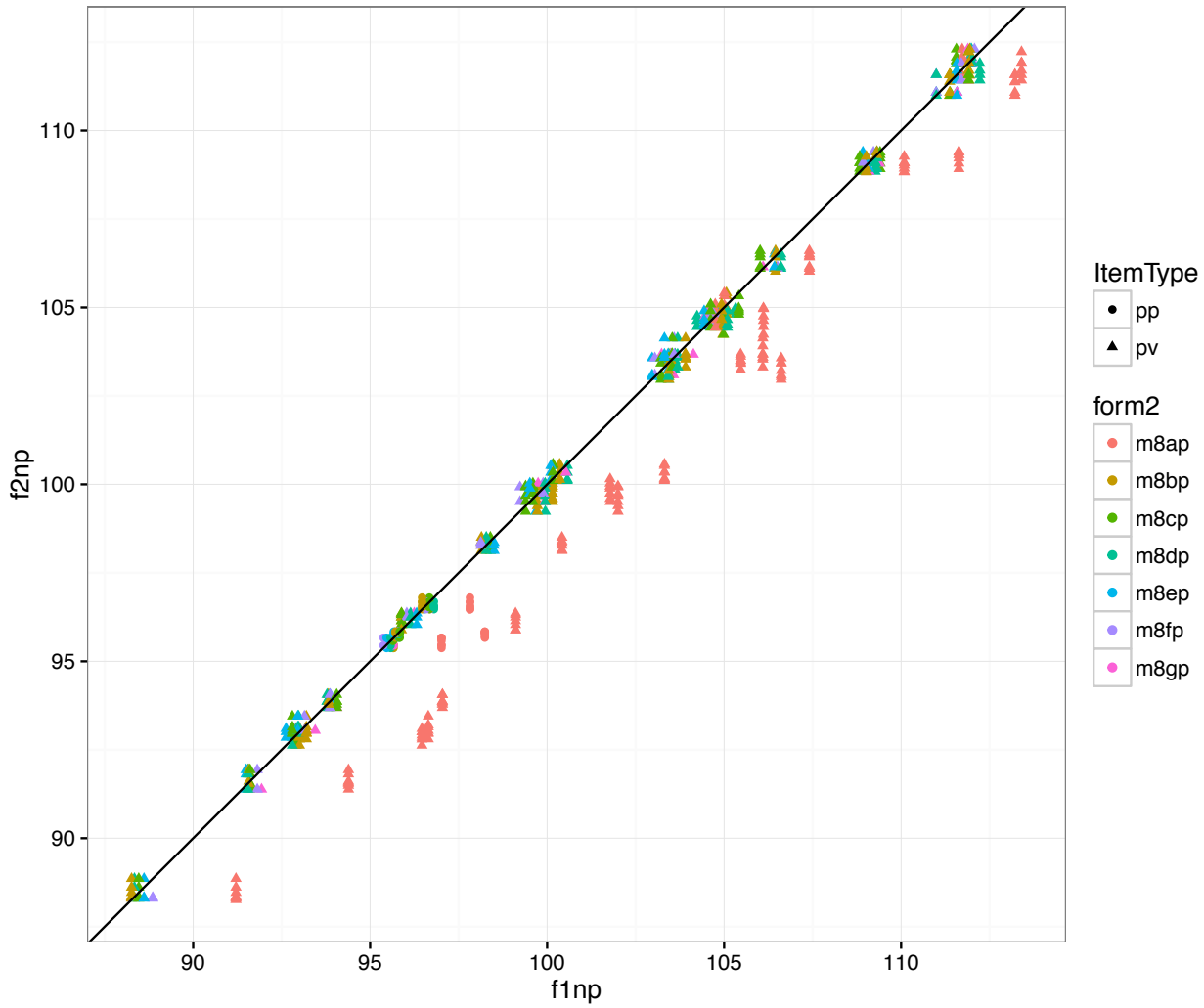


Figure 19.13: Math Grade 8 Item Stability Plot

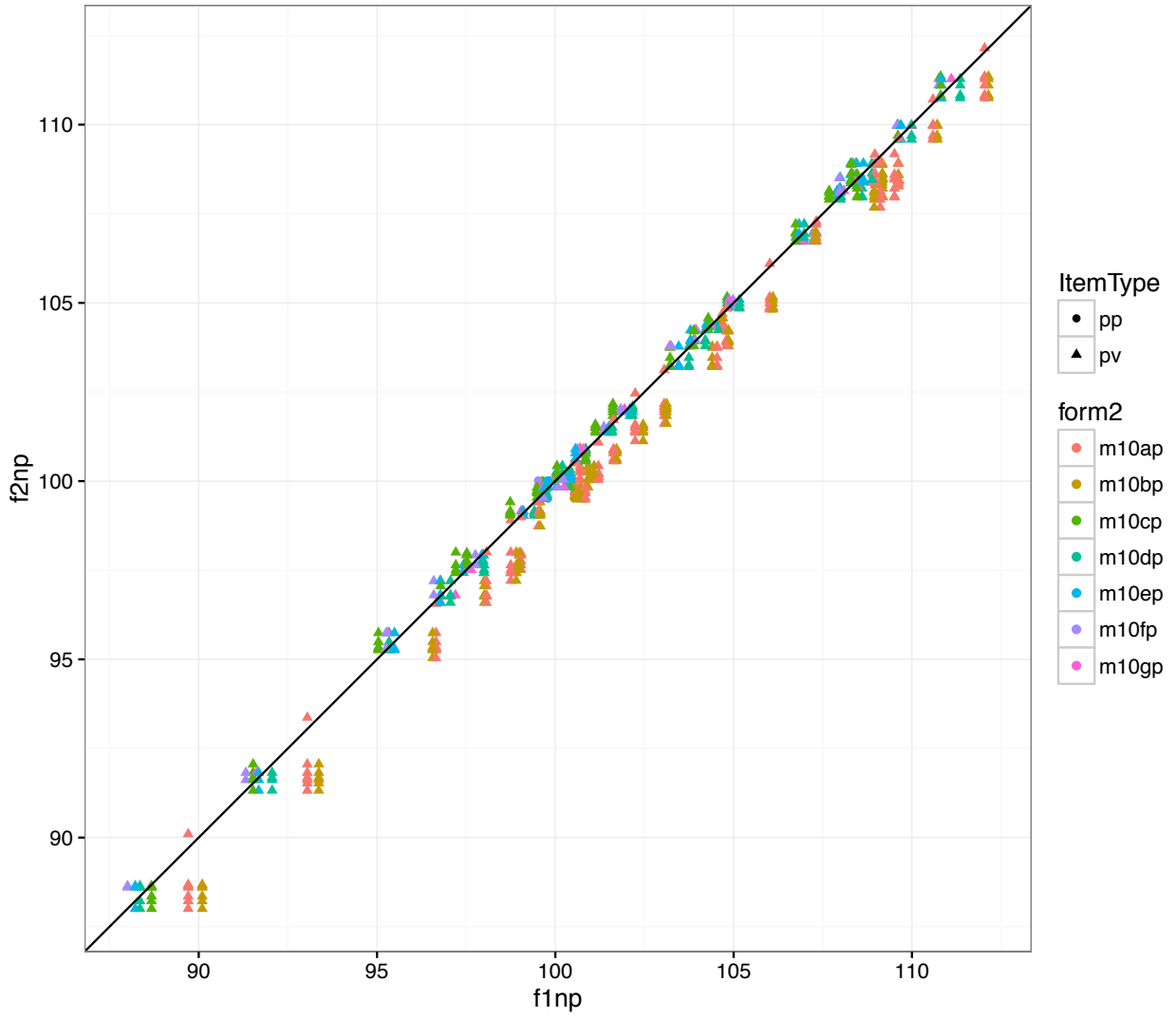


Figure 19.14: Math Grade 10 Item Stability Plot

19.3 Linking Procedures for Future KAP Assessments

In the future, a chained linking design will be utilized for the KAP operational tests. Each year, scores from the new KAP test forms will be linked to the scale of previous KAP test forms. The chain will originate from the scale defined during this year's test administration, which will serve as the reference for calibrating all future items in the item pool. When the item parameters from the new tests are placed on the bank's scale, the resulting scaled scores for the new test forms will be expressed on the scale constructed for this year's forms.

19.3.1 Preequating versus Postequating

When the transition to MST administrations begin, raw-score to scale-score conversion tables will be prepared in advance of the testing window. AAI will investigate if these preliminary tables can be used for official scoring of the KAP assessments. Such a procedure is known as *preequating*.

These preliminary tables are needed to appropriately route students across the MST stages.

AAI will compare the preequating results to results derived from a process called *postequating*. Postequating is considered the more conservative procedure as the functioning of some items can change between test administrations. If the changes are minimal in the first few years of the testing program, the preequating and postequating results may be very similar. In this case, preequating results might be preferred for score reporting in the future as it can lead to faster score reporting for parents and educators. Even if preequating is used, a rapid turn around postequating check on earlier testers is considered a prudent prophylactic measure.

19.3.2 Quality Control

In other testing programs, score linking has been an activity that has experienced human error. Consequently, quality control during the linking process is vital. AAI will ensure that all major linking analyses are independently run by two staff members. Their results will be compared and any differences resolved. Comparing results from different analyses can be helpful as well and there are usually several opportunities to cross-compare results. The n-counts and other item statistics should agree across different analyses.

Longitudinal trends of important statistics (e.g., average test scores, etc.) will be maintained for historical purposes. Although some professional judgment is needed regarding whether the statistical indices make sense or whether changes from a previous year are not reasonable, flags that arise during such comparisons are often the first

A sample historical statistics trend table is provided at the end of the operational statistics chapter.

indication that an error has occurred.

Program outputs will be inspected for any indication that there were any atypical endings to the program runs. The stability of anchor items between forms and across years will be checked using scatter plots (as done earlier in this chapter). A strong linear fit should result. The fit line should have a slope close to one and an intercept close to zero. Items straying substantially from the fit line indicate that they are not performing in the same way across forms or across years. These items must be reviewed to see if they have changed from their past usage to their current usage (wording, item position, etc.).

20
Reliability

This chapter addresses the reliability and precision of KAP test scores. Reliability refers to:

> the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker; the degree to which scores are free of random errors of measurement for a given group[1]

[1] From *Standards for Educational and Psychological Testing*, AERA, APA, & NCME, 2014, p. 222–223.

Reliability is intended to reflect the degree of inconsistency in test scores due to *random* error sources. Multiple sources of random error exist, yet most reliability indices only reflect a single source of error. Test users should try to understand what type of error is being considered, and even more importantly, what types are not. Systematic error sources also exist and can artificially increase reliability.

Some random error sources include the occasion of testing, the items selected, and if used, any raters who score the items.

Test length and population heterogeneity must be considered when evaluating reliability. Test scores based on more items tend to be more reliable than test scores based on fewer items. Additionally, reliability coefficients are group specific. Reliabilities tend to be higher in score populations that are more heterogeneous and lower in score populations that are more homogeneous.

Understanding the distinction between relative error and absolute error is also important. Many reliability indices only reflect relative error. Relative error is of interest when specific scores received do not matter, yet the relative ordering of students based on their test scores is of interest. When specific score values are important (e.g., if cut scores are used), then absolute error is important. There is more error variance when considering the absolute scores of examinees, which suggests lower reliability.

Systematic error will decrease the validity of any inferences made about test scores. The validity of KAP scores is further discussed in the next chapter.

It is often said that reliability concerns test scores and not the test specifically. However, test scores can be affected by characteristics of the testing instrument, such as the number of items.

The remainder of this chapter provides results and interpretations for the following:

- Reliability coefficients for total scores and claim scores

- Conditional standard errors of measurement (CSEMs)
- Decision consistency

20.1 Reliability Indices

The reliability coefficient expresses the consistency of test scores as the ratio of true-score variance to total-score variance. The total variance contains two components: (1) the variance in true scores—true individual differences in the attribute being measured, and (2) the variance from random fluctuations due to the imperfections in the measurement process (error variance).

While total variance equals true score variance plus error variance, a covariance term is not required in the denominator of the final formula, because true scores and errors are assumed to be uncorrelated in classical test theory.

$$\rho_X^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

Reliability coefficients range from 0.0 to 1.0. If all test score variances were true, the scores would be perfectly consistent and the index would equal 1.0. The index would be 0.0 if none of the test score variances were true. Such scores would be pure random noise (i.e., all measurement error).

Values of 1.0 are never achieved in practice. Larger coefficients are more desirable because they indicate that test scores are less influenced by random error.

Several different reliability indices exist. For tests scaled using item response theory (IRT), IRT marginal reliability is often reported. KAP marginal reliability estimates for each test form used at all grade levels are provided in the following tables.

Information about how IRT marginal reliability is computed is provided in the section on standard errors of measurement. As cautioned earlier, this index does not take into account other random sources of error (e.g., variations associated with the linking process; daily fluctuation in student health and behavior, the testing environment; rater inconsistency)

Subject	Grade	Form	Max. Score	Reliability
Math	3	A	64	0.91
Math	3	B	51	0.90
Math	3	C	63	0.91
Math	3	D	66	0.92
Math	3	E	52	0.90
Math	3	F	61	0.92
Math	3	G	64	0.92
Math	3	H	63	0.92

Table 20.1: Marginal Scaled Score Reliability for Grade 3 Math Summed Scores

All tabled values in this section are based on the full state population.

Subject	Grade	Form	Max. Score	Reliability
Math	4	A	67	0.92
Math	4	B	46	0.90
Math	4	C	67	0.92
Math	4	D	66	0.93
Math	4	E	51	0.91
Math	4	F	68	0.93
Math	4	G	67	0.93
Math	4	H	64	0.93

Table 20.2: Marginal Scaled Score Reliability for Grade 4 Math Summed Scores

Subject	Grade	Form	Max. Score	Reliability
Math	5	A	68	0.93
Math	5	B	52	0.90
Math	5	C	65	0.92
Math	5	D	65	0.92
Math	5	E	51	0.91
Math	5	F	62	0.92
Math	5	G	64	0.92
Math	5	H	65	0.92

Table 20.3: Marginal Scaled Score Reliability for Grade 5 Math Summed Scores

Subject	Grade	Form	Max. Score	Reliability
Math	6	A	69	0.90
Math	6	B	50	0.88
Math	6	C	64	0.90
Math	6	D	64	0.91
Math	6	E	49	0.89
Math	6	F	58	0.89
Math	6	G	65	0.92
Math	6	H	53	0.89

Table 20.4: Marginal Scaled Score Reliability for Grade 6 Math Summed Scores

Subject	Grade	Form	Max. Score	Reliability
Math	7	A	60	0.90
Math	7	B	47	0.88
Math	7	C	63	0.89
Math	7	D	62	0.90
Math	7	E	51	0.88
Math	7	F	65	0.90
Math	7	G	61	0.90
Math	7	H	48	0.89

Table 20.5: Marginal Scaled Score Reliability for Grade 7 Math Summed Scores

Subject	Grade	Form	Max. Score	Reliability
Math	8	A	63	0.89
Math	8	B	49	0.87
Math	8	C	56	0.88
Math	8	D	61	0.90
Math	8	E	48	0.87
Math	8	F	59	0.89
Math	8	G	61	0.90
Math	8	H	47	0.88

Table 20.6: Marginal Scaled Score Reliability for Grade 8 Math Summed Scores

Subject	Grade	Form	Max. Score	Reliability
Math	10	A	57	0.88
Math	10	B	55	0.88
Math	10	C	59	0.89
Math	10	D	58	0.88
Math	10	E	57	0.89
Math	10	F	56	0.89
Math	10	G	60	0.88
Math	10	H	52	0.88

Table 20.7: Marginal Scaled Score Reliability for Grade 10 Math Summed Scores

Subject	Grade	Form	Max. Score	Reliability
ELA	3	A	74	0.91
ELA	3	B	58	0.90
ELA	3	C	73	0.90
ELA	3	D	58	0.89
ELA	3	E	77	0.91
ELA	3	F	76	0.91
ELA	3	G	76	0.91
ELA	3	H	74	0.90

Table 20.8: Marginal Scaled Score Reliability for Grade 3 ELA Summed Scores

Subject	Grade	Form	Max. Score	Reliability
ELA	4	A	80	0.91
ELA	4	B	58	0.88
ELA	4	C	71	0.89
ELA	4	D	60	0.86
ELA	4	E	74	0.89
ELA	4	F	77	0.90
ELA	4	G	72	0.89
ELA	4	H	69	0.90

Table 20.9: Marginal Scaled Score Reliability for Grade 4 ELA Summed Scores

Subject	Grade	Form	Max. Score	Reliability
ELA	5	A	76	0.87
ELA	5	B	61	0.86
ELA	5	C	74	0.88
ELA	5	D	61	0.83
ELA	5	E	77	0.89
ELA	5	F	71	0.89
ELA	5	G	77	0.88
ELA	5	H	75	0.89

Table 20.10: Marginal Scaled Score Reliability for Grade 5 ELA Summed Scores

Subject	Grade	Form	Max. Score	Reliability
ELA	6	A	73	0.90
ELA	6	B	57	0.87
ELA	6	C	69	0.90
ELA	6	D	58	0.86
ELA	6	E	73	0.89
ELA	6	F	72	0.89
ELA	6	G	69	0.89
ELA	6	H	66	0.89

Table 20.11: Marginal Scaled Score Reliability for Grade 6 ELA Summed Scores

Subject	Grade	Form	Max. Score	Reliability
ELA	7	A	72	0.90
ELA	7	B	58	0.87
ELA	7	C	74	0.90
ELA	7	D	58	0.83
ELA	7	E	80	0.89
ELA	7	F	82	0.88
ELA	7	G	80	0.88
ELA	7	H	76	0.88

Table 20.12: Marginal Scaled Score Reliability for Grade 7 ELA Summed Scores

Subject	Grade	Form	Max. Score	Reliability
ELA	8	A	75	0.90
ELA	8	B	57	0.88
ELA	8	C	68	0.89
ELA	8	D	55	0.87
ELA	8	E	64	0.87
ELA	8	F	76	0.89
ELA	8	G	73	0.90
ELA	8	H	70	0.88

Table 20.13: Marginal Scaled Score Reliability for Grade 8 ELA Summed Scores

Subject	Grade	Form	Max. Score	Reliability
ELA	10	A	71	0.90
ELA	10	B	74	0.90
ELA	10	C	72	0.89
ELA	10	D	85	0.90
ELA	10	E	71	0.89
ELA	10	F	82	0.89
ELA	10	G	83	0.91
ELA	10	H	83	0.90

Table 20.14: Marginal Scaled Score Reliability for Grade 10 ELA Summed Scores

20.1.1 Interpretation Considerations

20.1.1.1 Rules of Thumb

The lower that a reliability coefficient is, the greater the potential for over-interpretation of the results. Yet, what reliability value is high enough? Although frequently requested, rules of thumb for interpreting the magnitude of reliability are arbitrary. On the low end, an informative point of reference is a reliability coefficient of 0.50. This represents the point where there is as much error variance as true-score variance in the scores.

A value of 0.80 is often heard as a threshold for interpreting individual scores.

A better approach is to research the reliabilities from similar testing instruments to determine what values are common. For KAP, comparisons to tests of similar lengths, administered to similar student populations, in other large-scale assessment programs would be relevant. The highest reliability values observed in other state assessment programs are usually in the low 0.90s, and values in the mid 0.80s are very common.

In the prior tables, form reliabilities tended to be higher when the forms had more points. Regardless of subject, grade, or form, all KAP total test score reliability values were adequate because all ranged from the low 0.80s to low 0.90s.

20.1.1.2 Biases Leading to Underestimates of Reliability

Some factors can negatively bias reliability, making it appear lower than it really is. One situation is when tests include a planned stratification of the test items according to specific content topics. Violation of the IRT model's unidimensionality assumption might occur and affect the marginal reliability estimates. Although KAP tests are built using content specifications, all total test score marginal reliabilities were above 0.80, indicating a considerable degree of consistency in the KAP test scores.

20.1.1.3 Biases Leading to Overestimates of Reliability

Any reliability index can understate the problem of measurement error when it does not account for alternate sources of random error that are nontrivial. Another positive bias can occur when items are associated (clustered) with a common stimulus. These cases can violate the IRT model's local item independence assumption and affect the marginal reliability estimates. Such a situation does not guarantee that the reliability estimate will be markedly affected, but the potential exists.

IRT marginal reliabilities do not take into account random sources of error such as those associated with the linking process and day-to-day variations in students (e.g., health, energy) and the testing environment.

Frequently used terms for this situation are *item bundles* and *testlets*. One concrete example is when multiple reading comprehension items are associated with a common passage selection.

Finally, all reliabilities reported in this chapter have a slight upward bias because they were based on the original raw scores which allowed

decimals. Before reporting, the decimal raw scores were rounded to the nearest integer, with decimal values of 0.5 rounded up. The magnitude in the loss of precision in going from raw scores with decimals to integer raw scores is difficult to quantify formally, but is believed to be a small, perhaps reducing reliability no more than 0.01 or 0.02.

This is based on the degree of attenuation often observed when going from IRT pattern scores to IRT summed scores.

20.2 Subgroups

TOTAL TEST SCORE reliabilities for important subgroups are documented in the appendix. Like the prior total test reliabilities for the full state population, the subgroup reliabilities are based on IRT marginal reliability estimates. Subgroup reliabilities ranged from the mid 80's to lower 90's.

Results are provided for all groups that had 100 or more students.

20.3 Claim Scores

AS NOTED EARLIER, higher reliabilities are associated with increased test length, and lower reliabilities are associated with decreased test length. The following figure illustrates this relationship for a hypothetical 60-point test with four total score reliabilities: 0.95 (red line), 0.90 (green line), 0.85 (blue line), and 0.80 (purple line). The green curve for reliability equal to 0.90 suggests that a 10-item claim score would be expected to have a score reliability of 0.60.

There are diminishing returns when adding test items. An increase from 10 items to 20 items adds more score consistency than an increase from 20 items to 30 items, and so on. Similarly, the loss in score consistency compounds exponentially when reducing test items.

These curves were projected using the Spearman-Brown prophecy formula. This formula assumes all items are exchangeable, which in practice they likely will not be. While this figure will not perfectly model actual claim score reliabilities, it illustrates the impact that limited numbers of items can have on reliability. In the following results, claim scores with more points tend to show higher reliability coefficients, and those with fewer points tend to show lower reliability coefficients. At some point, claim score reliabilities may be too low to warrant interpretation at the individual student level.

When test items cover specific claims, the claim items will likely be more homogeneous than items comprising the entire test. Consequently, the provided curves will underpredict the reliability of the claims.

20.3.1 Results

Claim score reliabilities are reported in the appendix. As expected, claim scores with more items tended to have higher reliability coefficients. However, the most significant result pertains to the fact that some claim score reliabilities are too low to warrant interpretation at the individual student level. Although there is no firm guideline regarding how low is too low for a reliability coefficient, many claim score reliabilities are below 0.50, which suggests that there is as much error variance as true-score variance in the scores. The lower the value

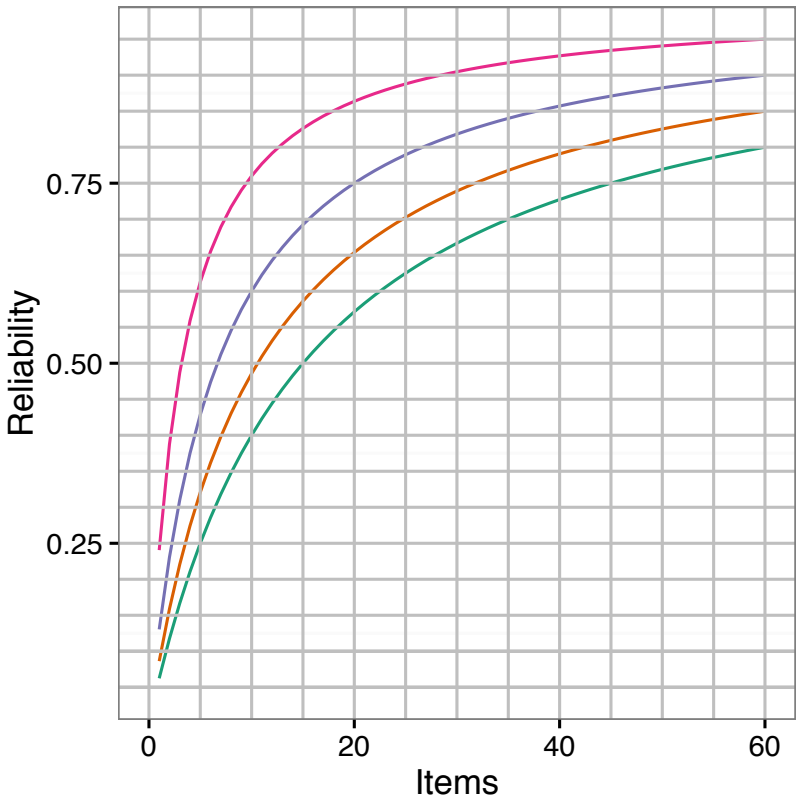


Figure 20.1: The Relationship Between Test Length and Test Score Reliability

of a given reliability coefficient, the greater the potential for over-interpretation of the score. For this reason, individual student reports for math do not report separate results for mathematics Claims 2, 3, or 4. Instead, these claim scores are combined on the student reports.

20.3.2 Group-Level Scores

The results presented in this chapter pertain to the reliability of individual scores. Group results (e.g., school, district, and state levels) are also provided on KAP score reports, but the reliability of those scores is not specifically calculated here. As a general rule, the reliabilities of group mean scores are higher (sometimes substantially so) than the corresponding reliabilities for individual scores. This is especially important to remember for claim scores because those scores can be quite reliable at the group level, even though their individual reliabilities may be too low to warrant interpretation.

Because the reliability of group mean scores (e.g., school or district means) tends to be higher than the reliability of individual scores, the interpretation of claim scores at these aggregate levels should be reasonable. Consequently, math Claims 2, 3, and 4 are reported separately at aggregate levels.

Even though the reliability of mean scores based only on a few items might be adequate, the validity of those same scores might be suspect because those few items may not adequately cover the construct of interest. Validity is further discussed in the next chapter.

20.4 Standard Error of Measurement

THE RELIABILITY COEFFICIENT is a unit-free indicator that reflects the degree to which scores are free of measurement error. It always ranges between 0.0 and 1.0 regardless of the test's scale. Reliability coefficients best reflect the extent to which measurement inconsistencies may be present or absent in a group. Unfortunately, they do not adequately support test score users who must interpret test results. In contrast, the conditional standard error of measurement (CSEM), another indicator of test score precision, is better suited for illustrating the effect of measurement inconsistencies on student test scores.

While a precise, theoretical interpretation of the SEM is somewhat unwieldy, the following scenario provides a beginning point for understanding the concept. If everyone being tested had the same true score, there would still be some variation in observed scores due to the aforementioned random imperfections in the measurement process. The standard error is defined as the standard deviation of the distribution of observed scores for students with identical true scores. Because the SEM is an index of the random variability in test scores in actual score units, it is a very important piece of information for test score users.

The *true score* is the score the person would receive if the measurement process were perfect.
The *standard deviation* of a distribution is a measure of the dispersion of the observations. For the normal distribution, about 32 percent of observations are more than one standard deviation above or below the mean.

The above description is rarely provided to test score users. A frequent description that is inaccurate in most instances is that the SEM represents a likely score range that might occur if a student could be tested twice with the same instrument. Unfortunately, this explanation implies that the only source of random error being considered is related to the occasion of testing.

The test-retest reliability coefficient captures random error associated with the occasion of testing. This is not the type of reliability computed for the KAP.

20.4.1 IRT Conditional Standard Error of Measurement

The CSEM indicates the degree of measurement error in scaled-score units. It varies as a function of actual scaled scores; and therefore, may be especially useful in characterizing measurement precision around a score level used for decision making, such as a cut score used for identifying students who meet a given performance standard.

When an IRT model is applied, the CSEM at any given point on the ability continuum is defined as the reciprocal of the square root of the test information function derived from the IRT scaling model. In the formula, $CSEM(\hat{\theta})$ is the conditional standard error of measurement, and $I(\hat{\theta})$ is the test information function. Test information depends on the sum of the corresponding information functions for the test items. Item information depends on each item's unique conditional item score variance as determined from its slope and threshold parameters.

$$CSEM(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}}$$

Figure 20.2: CSEM Formula on Theta Metric

The CSEM formula is presented using the IRT ability (θ) metric. The conditional standard error on the scaled-score (SS) metric is determined by multiplying the $CSEM(\hat{\theta})$ by the *slope* (or *multiplicative constant*, b,) of the linear transformation equation used to convert the IRT ability estimates to scaled scores.

$$CSEM(SS) = CSEM(\hat{\theta}) \times b$$

Figure 20.3: Converting the CSEM to the Scaled Score Metric

The linear transformation formulas for each KAP test are provided in the scaling chapter.

20.4.2 CSEM's Connection to Marginal Reliability

The marginal reliabilities reported earlier are actually related to the square of the CSEMs. Specifically, the integration (or averaging) over all θ values provides an average error variance for the test. Once $\bar{\sigma}^2_{e^*}$ is calculated, the marginal test reliability can be determined. Note that the formula denotes marginal reliability as $\bar{\rho}$ to explicitly indicate that it is an average.

$$\bar{\sigma}^2_{e^*} = \int \sigma^2_{e^*} g(\theta) d\theta$$

Figure 20.4: Average Error Variance Formula

$$\bar{\rho} = \frac{\sigma^2_{\theta} - \sigma^2_{e^*}}{\sigma^2_{\theta}}$$

Figure 20.5: Marginal Reliability Formula

20.4.3 Confidence Intervals (CIs)

CSEMs also allow statements regarding the precision of individual test scores by helping derive reasonable limits around observed scaled scores through construction of approximate score bands, referred to as *confidence intervals* (CIs). CIs are constructed by adding and subtracting a multiplicative factor of the CSEM from the observed scaled

score. Taking the $SS \pm CSEM \times 1.00$ constructs a 68% CI. Taking the $SS \pm CSEM \times 1.96$ constructs a 95% confidence interval.

68% CIs are used on KAP score reports.

Telling users that their CIs are based on the reciprocal of the square root of the test information function derived from the IRT scaling model is not helpful. A more pragmatic explanation of CIs, based on the error source considered by internal consistency reliability, is provided to KAP users on score reports. Specifically, users are told that the CIs provide reasonable bounds for the range of scores a student might receive if he or she took multiple, content-equivalent versions of the test (i.e., tests that covered exactly the same content but included different sets of items). As an example, if a student's KAP score was 350 and the 68% CI constant was 5, then the student would likely have received a score somewhere between 345 and 355 if he or she had taken a different version of test.

Because IRT CSEMs are based on statistical information, it is questionable whether they account for error variance due to items, except if the statistical properties of the items are exactly the same (Brennan, 2001). Therefore, it is difficult to construct a simple explanation of IRT CSEMs for the general public.

A more precise interpretation of the CI is that if a student were tested an infinite number of times, and 68% CIs were constructed for each of those scores, 68% of those CIs would capture the student's true score.

20.4.4 Results and Observations

The following figures document the IRT CSEMs associated with each scaled-score level by test form. As noted earlier, CSEMs change across the scaled-score range. However, for most CSEM plots, the values change slowly across a noticeably large range in the middle of the scaled scores, creating somewhat flat bottoms for the curves. The values increase at both extremes (i.e., at smaller and larger scaled scores), giving these figures their typical U-shaped pattern. The CSEM curves vary some across test forms but not too significantly.

CSEMs are not plotted for scaled scores more extreme than the lowest obtainable scaled scores (LOSS) and the highest obtainable scaled scores (HOSS) so the U-shape does not appear as pronounced as it might in most plots.

The three red, dashed lines represent the Level 2, 3, and 4 scaled score cuts, respectively, as values move from low to high. CSEM values at the cut scores were generally smaller, especially in mathematics, indicating more precise measurement occurs at these cuts. The curves for the ELA plots are shifted slightly to the left, making the Level 2 cut associated with the lowest CSEMs. However, given the slower changing CSEMs in the middle of the scale, the values at the Level 3 cut are still fairly small relative to the values at the extremes.

These CSEMs are for total test scores. Earlier in this chapter, total test reliabilities were reported by important subgroups. Assuming reasonable IRT model-data fit, the IRT-based CSEMs should not vary across groups. Model fit is explored in the item calibration chapter.

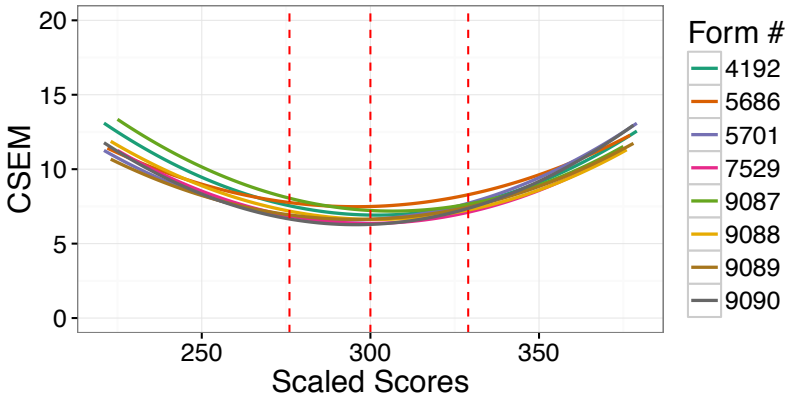


Figure 20.6: CSEMs for Grade 3 Math Summed Scores

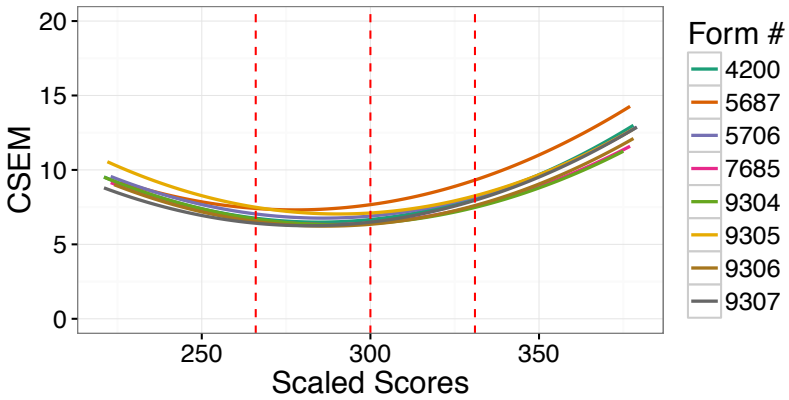


Figure 20.7: CSEMs for Grade 4 Math Summed Scores

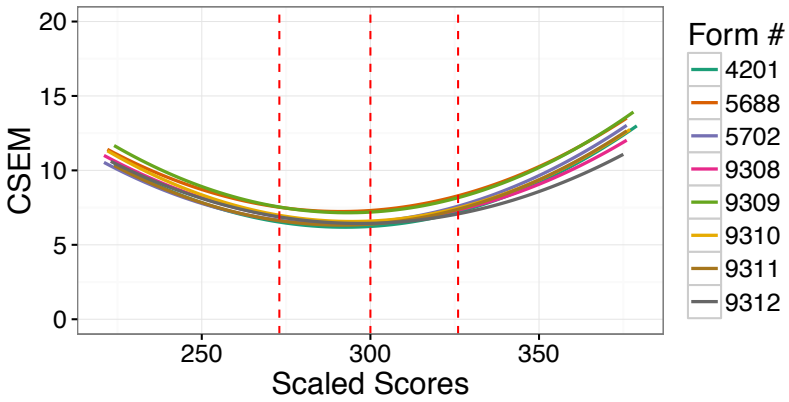


Figure 20.8: CSEMs for Grade 5 Math Summed Scores

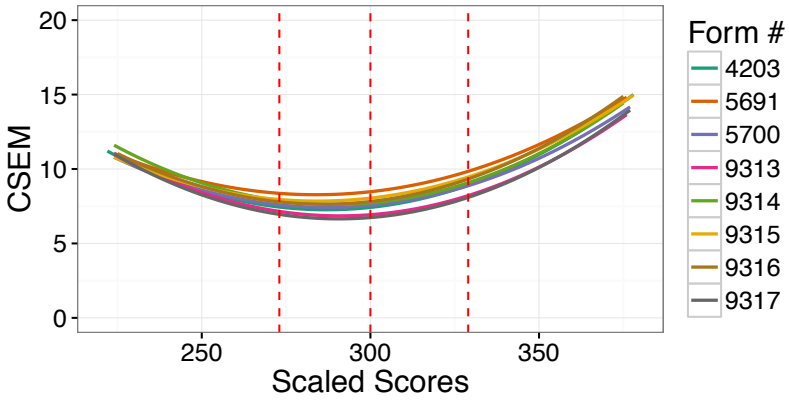


Figure 20.9: CSEMs for Grade 6 Math Summed Scores

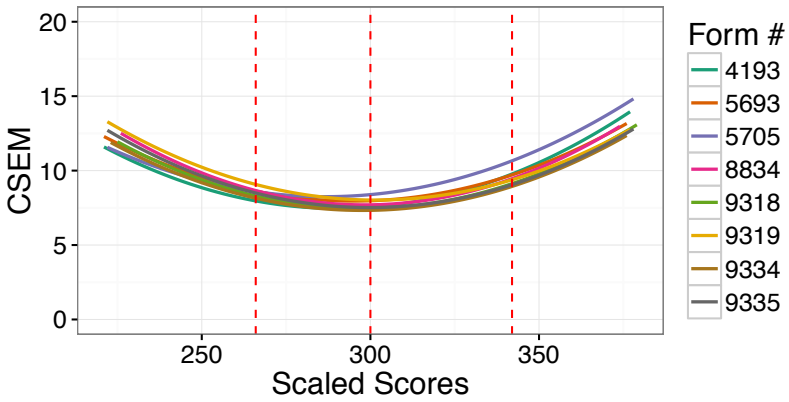


Figure 20.10: CSEMs for Grade 7 Math Summed Scores

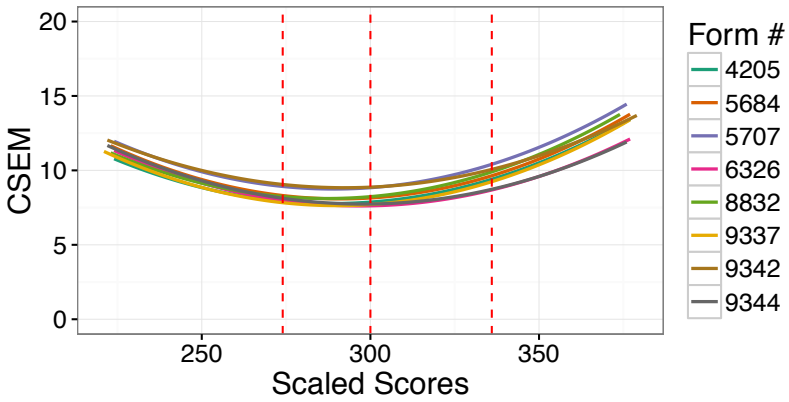


Figure 20.11: CSEMs for Grade 8 Math Summed Scores

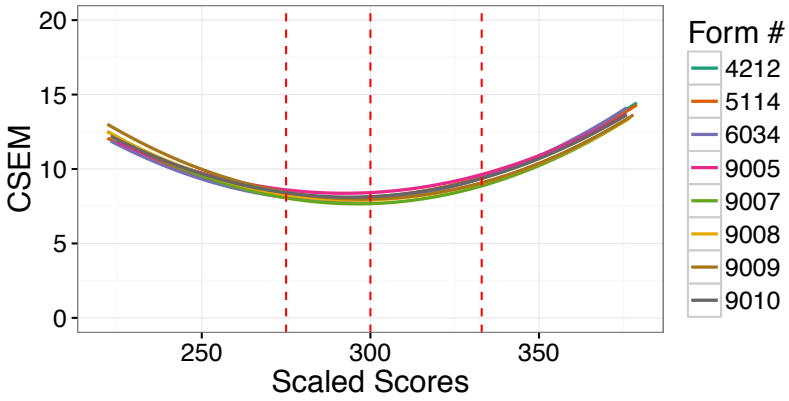


Figure 20.12: CSEMs for Grade 10 Math Summed Scores

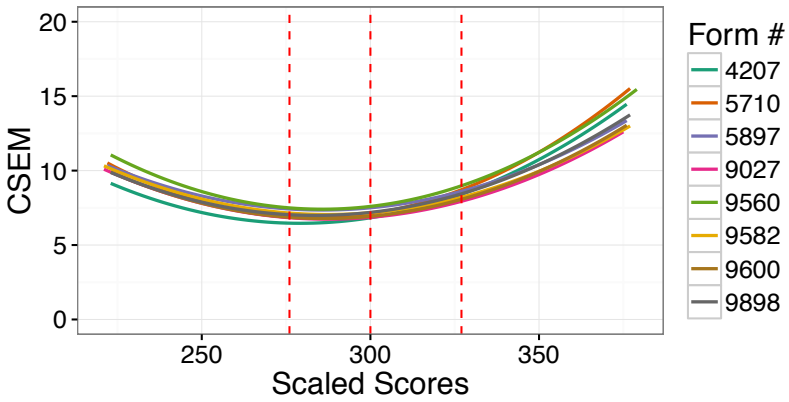


Figure 20.13: CSEMs for Grade 3 ELA Summed Scores

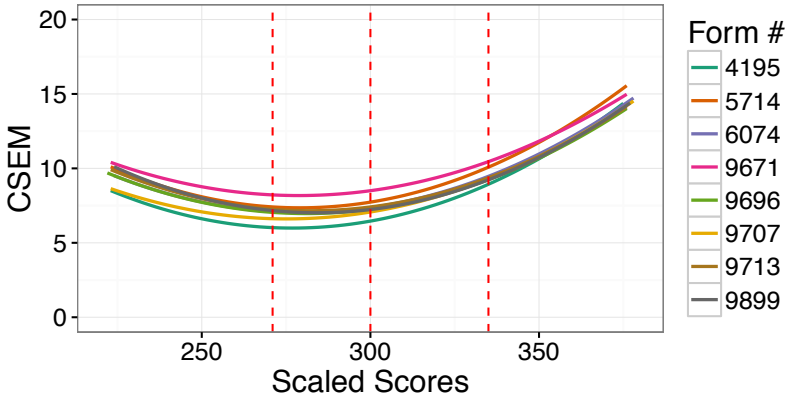


Figure 20.14: CSEMs for Grade 4 ELA Summed Scores

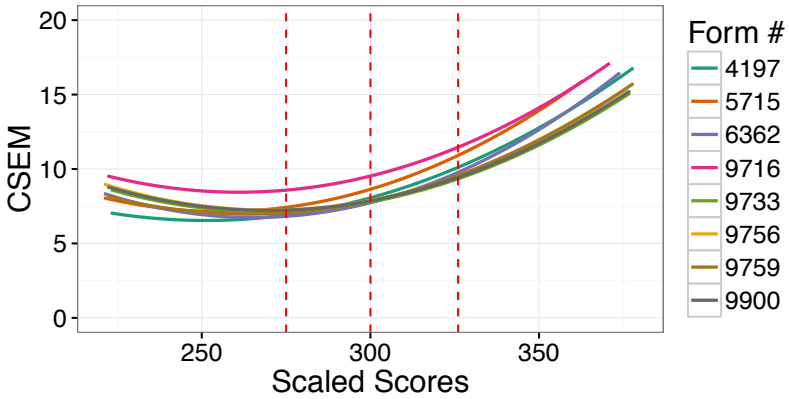


Figure 20.15: CSEMs for Grade 5 ELA Summed Scores

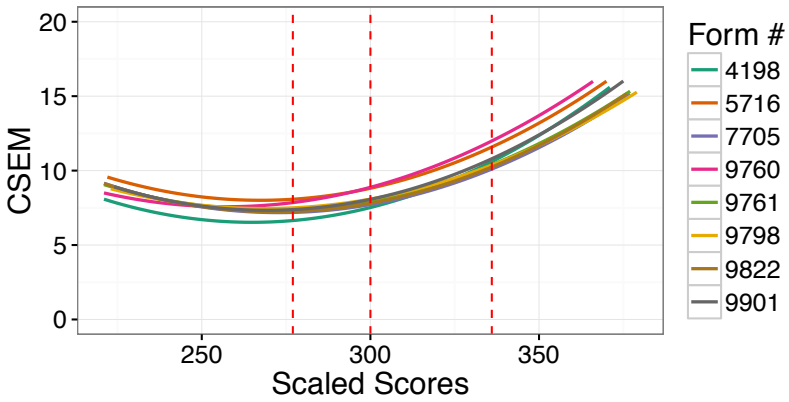


Figure 20.16: CSEMs for Grade 6 ELA Summed Scores

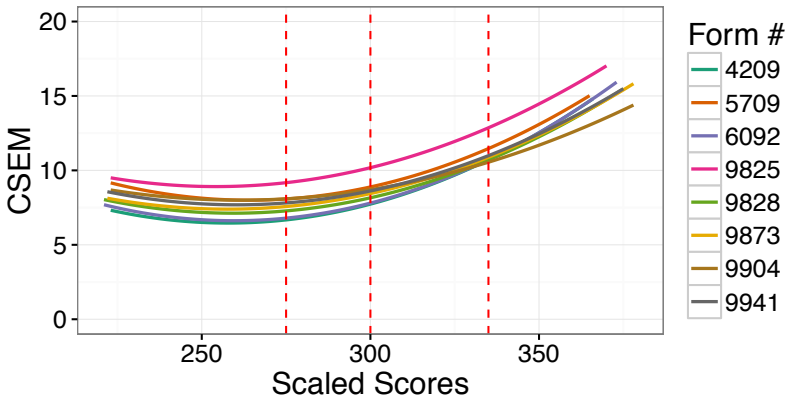


Figure 20.17: CSEMs for Grade 7 ELA Summed Scores

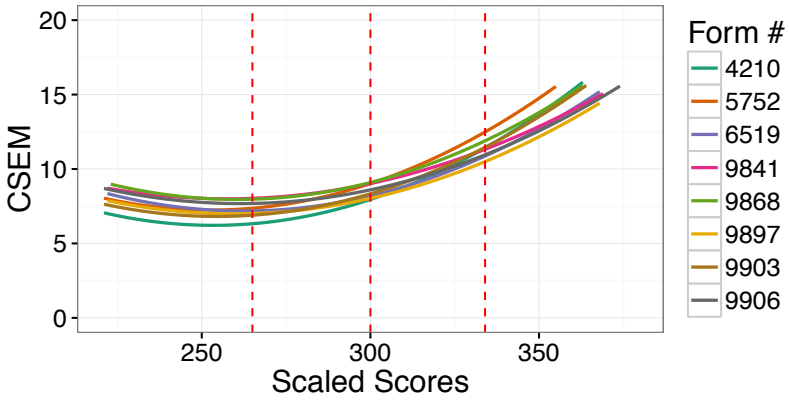


Figure 20.18: CSEMs for Grade 8 ELA Summed Scores

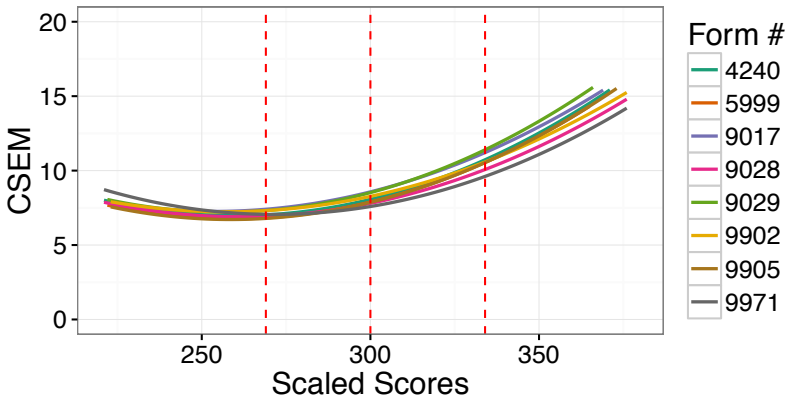


Figure 20.19: CSEMs for Grade 10 ELA Summed Scores

20.5 Decision Consistency and Accuracy

In accountability testing programs, there should be great interest in knowing how accurately students are classified into performance categories. If two parallel forms of the test were given to the same students, the consistency of measurement would be reflected by the extent that the classification decisions made from the first set of test scores matched the decisions based on the second set of test scores. *Classification consistency* refers to the degree to which the achievement level for each student can be replicated upon retesting using an equivalent form.[2] *Decision consistency* answers the question: What is the agreement between the classifications based on two non-overlapping, equally difficult forms of the test.

[2] Huynh (1976)

In other words, if a student is classified as being in one category based on Test One's score, how probable is it that the student would be classified in the same category if he or she took Test Two (a non-overlapping, equally difficult form of the test)? The proportions of correct decisions, ϕ, for two and four test score categories are computed by taking the sum of the diagonal entries from the tables below, that is, the proportion of students classified by the two forms into exactly the same achievement level. This proportion signifies the overall consistency.

	$Test\,1, Lvl\,1$	$Test\,1, Lvl\,2$	$Margin$
$Test\,2, Lvl\,1$	ϕ_{11}	ϕ_{12}	$\phi_{1\bullet}$
$Test\,2, Lvl\,2$	ϕ_{21}	ϕ_{22}	$\phi_{2\bullet}$
$Margin$	$\phi_{\bullet 1}$	$\phi_{\bullet 2}$	1

Figure 20.20: Pseudo-Decision Table for Two Hypothetical Categories

$$\phi = \phi_{11} + \phi_{22}$$

Figure 20.21: Proportion of Correct Decisions for a Two-Category Test

	$T1,\ L1$	$T1,\ L2$	$T1,\ L3$	$T1,\ L4$	$Margin$
$T2,\ L1$	ϕ_{11}	ϕ_{12}	ϕ_{13}	ϕ_{14}	$\phi_{1\bullet}$
$T2,\ L2$	ϕ_{21}	ϕ_{22}	ϕ_{23}	ϕ_{24}	$\phi_{2\bullet}$
$T2,\ L3$	ϕ_{31}	ϕ_{32}	ϕ_{33}	ϕ_{34}	$\phi_{3\bullet}$
$T2,\ L4$	ϕ_{41}	ϕ_{42}	ϕ_{43}	ϕ_{44}	$\phi_{4\bullet}$
$Margin$	$\phi_{\bullet 1}$	$\phi_{\bullet 2}$	$\phi_{\bullet 3}$	$\phi_{\bullet 4}$	1

Figure 20.22: Pseudo-Decision Table for Four Hypothetical Categories

$$\phi = \phi_{11} + \phi_{22} + \phi_{33} + \phi_{44}$$

Figure 20.23: Proportion of Correct Decisions for a Four-Category Test

Classification accuracy refers to the agreement of the observed classifications of students with the classifications made on the basis of their true scores. An observed score contains measurement error while

a true score is free of measurement error. A student's observed score can be formulated by the sum of his or her true score and measurement error. Decision accuracy is an index that determines the extent to which measurement error causes a classification different than expected from the true score.

True scores are unobserved and because it is not feasible to repeat KAP testing to estimate the proportion of students who would be classified in the same performance levels, a statistical model needs to be imposed on the data to estimate the true scores and to project the consistency and accuracy of classifications solely using data from the available administration.[3] Although a number of procedures are available, one well-known method[4] was developed utilizing a specific True Score Model. The results for overall consistency across all four performance levels as well as for the dichotomies created by the three cut scores are presented in the following tables. The tabled values are derived with the program BB-Class[5] using the Livingston and Lewis method.

[3] Hambleton and Novick (1973)

[4] Livingston and Lewis (1995)

[5] Brennan (2004)

The Livingston and Lewis approach is complex. Refer to the cited source for specific details regarding this approach.

Table 20.15: Decision Consistency and Accuracy for Grade 3 Math

Subject	Grade	Form	Cut	Accuracy	Consistency
Math	3	A	overall	0.76	0.67
Math	3	B	overall	0.77	0.68
Math	3	C	overall	0.79	0.70
Math	3	D	overall	0.79	0.71
Math	3	E	overall	0.77	0.68
Math	3	F	overall	0.79	0.70
Math	3	G	overall	0.78	0.70
Math	3	H	overall	0.79	0.71
Math	3	A	1 vs 2,3,4	0.88	0.83
Math	3	B	1 vs 2,3,4	0.95	0.93
Math	3	C	1 vs 2,3,4	0.95	0.93
Math	3	D	1 vs 2,3,4	0.94	0.92
Math	3	E	1 vs 2,3,4	0.94	0.92
Math	3	F	1 vs 2,3,4	0.94	0.92
Math	3	G	1 vs 2,3,4	0.94	0.92
Math	3	H	1 vs 2,3,4	0.95	0.93
Math	3	A	1,2 vs 3,4	0.92	0.88
Math	3	B	1,2 vs 3,4	0.90	0.85
Math	3	C	1,2 vs 3,4	0.90	0.86
Math	3	D	1,2 vs 3,4	0.91	0.87
Math	3	E	1,2 vs 3,4	0.90	0.86
Math	3	F	1,2 vs 3,4	0.90	0.86
Math	3	G	1,2 vs 3,4	0.90	0.86
Math	3	H	1,2 vs 3,4	0.91	0.87
Math	3	A	1,2,3 vs 4	0.96	0.95
Math	3	B	1,2,3 vs 4	0.93	0.90
Math	3	C	1,2,3 vs 4	0.93	0.91
Math	3	D	1,2,3 vs 4	0.94	0.91
Math	3	E	1,2,3 vs 4	0.93	0.90
Math	3	F	1,2,3 vs 4	0.94	0.92
Math	3	G	1,2,3 vs 4	0.94	0.91
Math	3	H	1,2,3 vs 4	0.94	0.91

Table 20.16: Decision Consistency and Accuracy for Grade 4 Math

Subject	Grade	Form	Cut	Accuracy	Consistency
Math	4	A	overall	0.81	0.73
Math	4	B	overall	0.80	0.72
Math	4	C	overall	0.81	0.74
Math	4	D	overall	0.82	0.75
Math	4	E	overall	0.81	0.73
Math	4	F	overall	0.82	0.75
Math	4	G	overall	0.83	0.75
Math	4	H	overall	0.82	0.74
Math	4	A	1 vs 2,3,4	0.89	0.85
Math	4	B	1 vs 2,3,4	0.95	0.92
Math	4	C	1 vs 2,3,4	0.95	0.92
Math	4	D	1 vs 2,3,4	0.95	0.93
Math	4	E	1 vs 2,3,4	0.95	0.93
Math	4	F	1 vs 2,3,4	0.95	0.93
Math	4	G	1 vs 2,3,4	0.95	0.93
Math	4	H	1 vs 2,3,4	0.95	0.93
Math	4	A	1,2 vs 3,4	0.94	0.91
Math	4	B	1,2 vs 3,4	0.89	0.85
Math	4	C	1,2 vs 3,4	0.91	0.87
Math	4	D	1,2 vs 3,4	0.91	0.88
Math	4	E	1,2 vs 3,4	0.90	0.86
Math	4	F	1,2 vs 3,4	0.91	0.88
Math	4	G	1,2 vs 3,4	0.91	0.88
Math	4	H	1,2 vs 3,4	0.91	0.87
Math	4	A	1,2,3 vs 4	0.98	0.97
Math	4	B	1,2,3 vs 4	0.96	0.94
Math	4	C	1,2,3 vs 4	0.96	0.94
Math	4	D	1,2,3 vs 4	0.96	0.95
Math	4	E	1,2,3 vs 4	0.96	0.94
Math	4	F	1,2,3 vs 4	0.96	0.94
Math	4	G	1,2,3 vs 4	0.96	0.95
Math	4	H	1,2,3 vs 4	0.96	0.94

Table 20.17: Decision Consistency and Accuracy for Grade 5 Math

Subject	Grade	Form	Cut	Accuracy	Consistency
Math	5	A	overall	0.82	0.74
Math	5	B	overall	0.77	0.68
Math	5	C	overall	0.79	0.71
Math	5	D	overall	0.79	0.71
Math	5	E	overall	0.78	0.69
Math	5	F	overall	0.79	0.70
Math	5	G	overall	0.80	0.71
Math	5	H	overall	0.78	0.70
Math	5	A	1 vs 2,3,4	0.89	0.84
Math	5	B	1 vs 2,3,4	0.91	0.88
Math	5	C	1 vs 2,3,4	0.92	0.89
Math	5	D	1 vs 2,3,4	0.92	0.89
Math	5	E	1 vs 2,3,4	0.92	0.89
Math	5	F	1 vs 2,3,4	0.91	0.88
Math	5	G	1 vs 2,3,4	0.92	0.89
Math	5	H	1 vs 2,3,4	0.92	0.89
Math	5	A	1,2 vs 3,4	0.95	0.93
Math	5	B	1,2 vs 3,4	0.90	0.87
Math	5	C	1,2 vs 3,4	0.92	0.88
Math	5	D	1,2 vs 3,4	0.91	0.88
Math	5	E	1,2 vs 3,4	0.90	0.86
Math	5	F	1,2 vs 3,4	0.92	0.88
Math	5	G	1,2 vs 3,4	0.91	0.88
Math	5	H	1,2 vs 3,4	0.91	0.88
Math	5	A	1,2,3 vs 4	0.98	0.97
Math	5	B	1,2,3 vs 4	0.95	0.93
Math	5	C	1,2,3 vs 4	0.96	0.94
Math	5	D	1,2,3 vs 4	0.96	0.94
Math	5	E	1,2,3 vs 4	0.95	0.93
Math	5	F	1,2,3 vs 4	0.96	0.94
Math	5	G	1,2,3 vs 4	0.96	0.94
Math	5	H	1,2,3 vs 4	0.95	0.94

Table 20.18: Decision Consistency and Accuracy for Grade 6 Math

Subject	Grade	Form	Cut	Accuracy	Consistency
Math	6	A	overall	0.78	0.69
Math	6	B	overall	0.75	0.65
Math	6	C	overall	0.76	0.66
Math	6	D	overall	0.78	0.69
Math	6	E	overall	0.77	0.67
Math	6	F	overall	0.76	0.67
Math	6	G	overall	0.79	0.70
Math	6	H	overall	0.76	0.67
Math	6	A	1 vs 2,3,4	0.86	0.81
Math	6	B	1 vs 2,3,4	0.89	0.84
Math	6	C	1 vs 2,3,4	0.89	0.85
Math	6	D	1 vs 2,3,4	0.91	0.87
Math	6	E	1 vs 2,3,4	0.91	0.87
Math	6	F	1 vs 2,3,4	0.91	0.87
Math	6	G	1 vs 2,3,4	0.91	0.88
Math	6	H	1 vs 2,3,4	0.91	0.87
Math	6	A	1,2 vs 3,4	0.94	0.92
Math	6	B	1,2 vs 3,4	0.90	0.86
Math	6	C	1,2 vs 3,4	0.91	0.87
Math	6	D	1,2 vs 3,4	0.91	0.88
Math	6	E	1,2 vs 3,4	0.90	0.87
Math	6	F	1,2 vs 3,4	0.90	0.86
Math	6	G	1,2 vs 3,4	0.92	0.88
Math	6	H	1,2 vs 3,4	0.90	0.86
Math	6	A	1,2,3 vs 4	0.98	0.97
Math	6	B	1,2,3 vs 4	0.96	0.94
Math	6	C	1,2,3 vs 4	0.96	0.94
Math	6	D	1,2,3 vs 4	0.96	0.95
Math	6	E	1,2,3 vs 4	0.96	0.94
Math	6	F	1,2,3 vs 4	0.95	0.94
Math	6	G	1,2,3 vs 4	0.96	0.94
Math	6	H	1,2,3 vs 4	0.96	0.94

Table 20.19: Decision Consistency and Accuracy for Grade 7 Math

Subject	Grade	Form	Cut	Accuracy	Consistency
Math	7	A	overall	0.78	0.69
Math	7	B	overall	0.80	0.72
Math	7	C	overall	0.80	0.72
Math	7	D	overall	0.82	0.74
Math	7	E	overall	0.80	0.72
Math	7	F	overall	0.80	0.72
Math	7	G	overall	0.80	0.72
Math	7	H	overall	0.80	0.72
Math	7	A	1 vs 2,3,4	0.85	0.79
Math	7	B	1 vs 2,3,4	0.92	0.89
Math	7	C	1 vs 2,3,4	0.92	0.88
Math	7	D	1 vs 2,3,4	0.92	0.89
Math	7	E	1 vs 2,3,4	0.92	0.89
Math	7	F	1 vs 2,3,4	0.90	0.86
Math	7	G	1 vs 2,3,4	0.90	0.87
Math	7	H	1 vs 2,3,4	0.91	0.87
Math	7	A	1,2 vs 3,4	0.94	0.92
Math	7	B	1,2 vs 3,4	0.90	0.86
Math	7	C	1,2 vs 3,4	0.91	0.87
Math	7	D	1,2 vs 3,4	0.91	0.88
Math	7	E	1,2 vs 3,4	0.90	0.86
Math	7	F	1,2 vs 3,4	0.92	0.89
Math	7	G	1,2 vs 3,4	0.92	0.88
Math	7	H	1,2 vs 3,4	0.91	0.87
Math	7	A	1,2,3 vs 4	0.99	0.98
Math	7	B	1,2,3 vs 4	0.98	0.97
Math	7	C	1,2,3 vs 4	0.98	0.97
Math	7	D	1,2,3 vs 4	0.98	0.98
Math	7	E	1,2,3 vs 4	0.98	0.97
Math	7	F	1,2,3 vs 4	0.98	0.97
Math	7	G	1,2,3 vs 4	0.98	0.97
Math	7	H	1,2,3 vs 4	0.98	0.97

Table 20.20: Decision Consistency and Accuracy for Grade 8 Math

Subject	Grade	Form	Cut	Accuracy	Consistency
Math	8	A	overall	0.82	0.74
Math	8	B	overall	0.75	0.65
Math	8	C	overall	0.76	0.66
Math	8	D	overall	0.77	0.68
Math	8	E	overall	0.74	0.65
Math	8	F	overall	0.76	0.67
Math	8	G	overall	0.78	0.69
Math	8	H	overall	0.77	0.68
Math	8	A	1 vs 2,3,4	0.88	0.83
Math	8	B	1 vs 2,3,4	0.85	0.80
Math	8	C	1 vs 2,3,4	0.86	0.81
Math	8	D	1 vs 2,3,4	0.87	0.82
Math	8	E	1 vs 2,3,4	0.86	0.80
Math	8	F	1 vs 2,3,4	0.86	0.81
Math	8	G	1 vs 2,3,4	0.87	0.82
Math	8	H	1 vs 2,3,4	0.88	0.84
Math	8	A	1,2 vs 3,4	0.95	0.93
Math	8	B	1,2 vs 3,4	0.92	0.89
Math	8	C	1,2 vs 3,4	0.92	0.89
Math	8	D	1,2 vs 3,4	0.93	0.90
Math	8	E	1,2 vs 3,4	0.91	0.88
Math	8	F	1,2 vs 3,4	0.92	0.89
Math	8	G	1,2 vs 3,4	0.93	0.90
Math	8	H	1,2 vs 3,4	0.91	0.88
Math	8	A	1,2,3 vs 4	0.99	0.98
Math	8	B	1,2,3 vs 4	0.98	0.97
Math	8	C	1,2,3 vs 4	0.98	0.97
Math	8	D	1,2,3 vs 4	0.98	0.97
Math	8	E	1,2,3 vs 4	0.97	0.96
Math	8	F	1,2,3 vs 4	0.98	0.97
Math	8	G	1,2,3 vs 4	0.98	0.97
Math	8	H	1,2,3 vs 4	0.98	0.97

Table 20.21: Decision Consistency and Accuracy for Grade 10 Math

Subject	Grade	Form	Cut	Accuracy	Consistency
Math	10	A	overall	0.79	0.71
Math	10	B	overall	0.78	0.69
Math	10	C	overall	0.77	0.68
Math	10	D	overall	0.76	0.66
Math	10	E	overall	0.77	0.67
Math	10	F	overall	0.76	0.67
Math	10	G	overall	0.75	0.65
Math	10	H	overall	0.77	0.68
Math	10	A	1 vs 2,3,4	0.88	0.83
Math	10	B	1 vs 2,3,4	0.87	0.81
Math	10	C	1 vs 2,3,4	0.88	0.83
Math	10	D	1 vs 2,3,4	0.85	0.80
Math	10	E	1 vs 2,3,4	0.87	0.81
Math	10	F	1 vs 2,3,4	0.86	0.81
Math	10	G	1 vs 2,3,4	0.85	0.79
Math	10	H	1 vs 2,3,4	0.87	0.82
Math	10	A	1,2 vs 3,4	0.93	0.90
Math	10	B	1,2 vs 3,4	0.94	0.91
Math	10	C	1,2 vs 3,4	0.92	0.89
Math	10	D	1,2 vs 3,4	0.93	0.90
Math	10	E	1,2 vs 3,4	0.93	0.90
Math	10	F	1,2 vs 3,4	0.93	0.89
Math	10	G	1,2 vs 3,4	0.93	0.90
Math	10	H	1,2 vs 3,4	0.92	0.89
Math	10	A	1,2,3 vs 4	0.98	0.97
Math	10	B	1,2,3 vs 4	0.98	0.97
Math	10	C	1,2,3 vs 4	0.97	0.96
Math	10	D	1,2,3 vs 4	0.98	0.97
Math	10	E	1,2,3 vs 4	0.97	0.96
Math	10	F	1,2,3 vs 4	0.98	0.97
Math	10	G	1,2,3 vs 4	0.98	0.97
Math	10	H	1,2,3 vs 4	0.98	0.97

Table 20.22: Decision Consistency and Accuracy for Grade 3 ELA

Subject	Grade	Form	Cut	Accuracy	Consistency
ELA	3	A	overall	0.78	0.70
ELA	3	B	overall	0.75	0.66
ELA	3	C	overall	0.76	0.66
ELA	3	D	overall	0.75	0.65
ELA	3	E	overall	0.77	0.68
ELA	3	F	overall	0.77	0.68
ELA	3	G	overall	0.77	0.67
ELA	3	H	overall	0.76	0.67
ELA	3	A	1 vs 2,3,4	0.91	0.87
ELA	3	B	1 vs 2,3,4	0.94	0.92
ELA	3	C	1 vs 2,3,4	0.94	0.91
ELA	3	D	1 vs 2,3,4	0.93	0.91
ELA	3	E	1 vs 2,3,4	0.94	0.92
ELA	3	F	1 vs 2,3,4	0.94	0.91
ELA	3	G	1 vs 2,3,4	0.94	0.92
ELA	3	H	1 vs 2,3,4	0.93	0.91
ELA	3	A	1,2 vs 3,4	0.92	0.89
ELA	3	B	1,2 vs 3,4	0.90	0.85
ELA	3	C	1,2 vs 3,4	0.89	0.85
ELA	3	D	1,2 vs 3,4	0.89	0.85
ELA	3	E	1,2 vs 3,4	0.90	0.86
ELA	3	F	1,2 vs 3,4	0.90	0.86
ELA	3	G	1,2 vs 3,4	0.90	0.86
ELA	3	H	1,2 vs 3,4	0.90	0.86
ELA	3	A	1,2,3 vs 4	0.96	0.94
ELA	3	B	1,2,3 vs 4	0.91	0.88
ELA	3	C	1,2,3 vs 4	0.92	0.89
ELA	3	D	1,2,3 vs 4	0.92	0.89
ELA	3	E	1,2,3 vs 4	0.93	0.91
ELA	3	F	1,2,3 vs 4	0.93	0.90
ELA	3	G	1,2,3 vs 4	0.92	0.89
ELA	3	H	1,2,3 vs 4	0.93	0.90

Table 20.23: Decision Consistency and Accuracy for Grade 4 ELA

Subject	Grade	Form	Cut	Accuracy	Consistency
ELA	4	A	overall	0.79	0.70
ELA	4	B	overall	0.77	0.68
ELA	4	C	overall	0.78	0.70
ELA	4	D	overall	0.76	0.67
ELA	4	E	overall	0.78	0.69
ELA	4	F	overall	0.77	0.68
ELA	4	G	overall	0.78	0.70
ELA	4	H	overall	0.79	0.71
ELA	4	A	1 vs 2,3,4	0.91	0.87
ELA	4	B	1 vs 2,3,4	0.97	0.95
ELA	4	C	1 vs 2,3,4	0.96	0.95
ELA	4	D	1 vs 2,3,4	0.96	0.95
ELA	4	E	1 vs 2,3,4	0.97	0.95
ELA	4	F	1 vs 2,3,4	0.97	0.96
ELA	4	G	1 vs 2,3,4	0.97	0.95
ELA	4	H	1 vs 2,3,4	0.97	0.96
ELA	4	A	1,2 vs 3,4	0.92	0.88
ELA	4	B	1,2 vs 3,4	0.89	0.85
ELA	4	C	1,2 vs 3,4	0.89	0.85
ELA	4	D	1,2 vs 3,4	0.89	0.84
ELA	4	E	1,2 vs 3,4	0.90	0.86
ELA	4	F	1,2 vs 3,4	0.90	0.86
ELA	4	G	1,2 vs 3,4	0.90	0.86
ELA	4	H	1,2 vs 3,4	0.90	0.86
ELA	4	A	1,2,3 vs 4	0.96	0.94
ELA	4	B	1,2,3 vs 4	0.91	0.88
ELA	4	C	1,2,3 vs 4	0.92	0.90
ELA	4	D	1,2,3 vs 4	0.91	0.88
ELA	4	E	1,2,3 vs 4	0.91	0.87
ELA	4	F	1,2,3 vs 4	0.90	0.86
ELA	4	G	1,2,3 vs 4	0.92	0.88
ELA	4	H	1,2,3 vs 4	0.92	0.89

Table 20.24: Decision Consistency and Accuracy for Grade 5 ELA

Subject	Grade	Form	Cut	Accuracy	Consistency
ELA	5	A	overall	0.72	0.63
ELA	5	B	overall	0.68	0.59
ELA	5	C	overall	0.72	0.62
ELA	5	D	overall	0.66	0.56
ELA	5	E	overall	0.74	0.64
ELA	5	F	overall	0.74	0.64
ELA	5	G	overall	0.72	0.62
ELA	5	H	overall	0.74	0.64
ELA	5	A	1 vs 2,3,4	0.90	0.86
ELA	5	B	1 vs 2,3,4	0.95	0.92
ELA	5	C	1 vs 2,3,4	0.95	0.93
ELA	5	D	1 vs 2,3,4	0.94	0.92
ELA	5	E	1 vs 2,3,4	0.95	0.93
ELA	5	F	1 vs 2,3,4	0.95	0.93
ELA	5	G	1 vs 2,3,4	0.95	0.93
ELA	5	H	1 vs 2,3,4	0.95	0.92
ELA	5	A	1,2 vs 3,4	0.89	0.85
ELA	5	B	1,2 vs 3,4	0.88	0.82
ELA	5	C	1,2 vs 3,4	0.89	0.84
ELA	5	D	1,2 vs 3,4	0.87	0.81
ELA	5	E	1,2 vs 3,4	0.89	0.84
ELA	5	F	1,2 vs 3,4	0.89	0.84
ELA	5	G	1,2 vs 3,4	0.89	0.84
ELA	5	H	1,2 vs 3,4	0.89	0.84
ELA	5	A	1,2,3 vs 4	0.93	0.90
ELA	5	B	1,2,3 vs 4	0.85	0.82
ELA	5	C	1,2,3 vs 4	0.88	0.84
ELA	5	D	1,2,3 vs 4	0.85	0.81
ELA	5	E	1,2,3 vs 4	0.90	0.86
ELA	5	F	1,2,3 vs 4	0.91	0.87
ELA	5	G	1,2,3 vs 4	0.88	0.84
ELA	5	H	1,2,3 vs 4	0.90	0.86

Table 20.25: Decision Consistency and Accuracy for Grade 6 ELA

Subject	Grade	Form	Cut	Accuracy	Consistency
ELA	6	A	overall	0.80	0.72
ELA	6	B	overall	0.75	0.66
ELA	6	C	overall	0.77	0.68
ELA	6	D	overall	0.76	0.66
ELA	6	E	overall	0.77	0.69
ELA	6	F	overall	0.77	0.68
ELA	6	G	overall	0.76	0.67
ELA	6	H	overall	0.76	0.68
ELA	6	A	1 vs 2,3,4	0.91	0.87
ELA	6	B	1 vs 2,3,4	0.91	0.87
ELA	6	C	1 vs 2,3,4	0.93	0.90
ELA	6	D	1 vs 2,3,4	0.92	0.89
ELA	6	E	1 vs 2,3,4	0.93	0.90
ELA	6	F	1 vs 2,3,4	0.93	0.90
ELA	6	G	1 vs 2,3,4	0.92	0.89
ELA	6	H	1 vs 2,3,4	0.92	0.89
ELA	6	A	1,2 vs 3,4	0.92	0.88
ELA	6	B	1,2 vs 3,4	0.87	0.83
ELA	6	C	1,2 vs 3,4	0.89	0.85
ELA	6	D	1,2 vs 3,4	0.87	0.82
ELA	6	E	1,2 vs 3,4	0.89	0.84
ELA	6	F	1,2 vs 3,4	0.89	0.84
ELA	6	G	1,2 vs 3,4	0.89	0.84
ELA	6	H	1,2 vs 3,4	0.88	0.84
ELA	6	A	1,2,3 vs 4	0.97	0.96
ELA	6	B	1,2,3 vs 4	0.96	0.95
ELA	6	C	1,2,3 vs 4	0.95	0.94
ELA	6	D	1,2,3 vs 4	0.97	0.95
ELA	6	E	1,2,3 vs 4	0.96	0.95
ELA	6	F	1,2,3 vs 4	0.96	0.94
ELA	6	G	1,2,3 vs 4	0.95	0.93
ELA	6	H	1,2,3 vs 4	0.96	0.94

Table 20.26: Decision Consistency and Accuracy for Grade 7 ELA

Subject	Grade	Form	Cut	Accuracy	Consistency
ELA	7	A	overall	0.80	0.72
ELA	7	B	overall	0.77	0.67
ELA	7	C	overall	0.76	0.67
ELA	7	D	overall	0.73	0.63
ELA	7	E	overall	0.78	0.69
ELA	7	F	overall	0.77	0.68
ELA	7	G	overall	0.77	0.68
ELA	7	H	overall	0.77	0.68
ELA	7	A	1 vs 2,3,4	0.91	0.87
ELA	7	B	1 vs 2,3,4	0.93	0.90
ELA	7	C	1 vs 2,3,4	0.93	0.90
ELA	7	D	1 vs 2,3,4	0.92	0.89
ELA	7	E	1 vs 2,3,4	0.93	0.90
ELA	7	F	1 vs 2,3,4	0.92	0.89
ELA	7	G	1 vs 2,3,4	0.93	0.90
ELA	7	H	1 vs 2,3,4	0.93	0.90
ELA	7	A	1,2 vs 3,4	0.91	0.87
ELA	7	B	1,2 vs 3,4	0.87	0.82
ELA	7	C	1,2 vs 3,4	0.89	0.84
ELA	7	D	1,2 vs 3,4	0.85	0.79
ELA	7	E	1,2 vs 3,4	0.88	0.83
ELA	7	F	1,2 vs 3,4	0.88	0.83
ELA	7	G	1,2 vs 3,4	0.88	0.83
ELA	7	H	1,2 vs 3,4	0.88	0.83
ELA	7	A	1,2,3 vs 4	0.98	0.97
ELA	7	B	1,2,3 vs 4	0.97	0.95
ELA	7	C	1,2,3 vs 4	0.94	0.93
ELA	7	D	1,2,3 vs 4	0.96	0.94
ELA	7	E	1,2,3 vs 4	0.96	0.95
ELA	7	F	1,2,3 vs 4	0.97	0.96
ELA	7	G	1,2,3 vs 4	0.96	0.95
ELA	7	H	1,2,3 vs 4	0.96	0.94

Table 20.27: Decision Consistency and Accuracy for Grade 8 ELA

Subject	Grade	Form	Cut	Accuracy	Consistency
ELA	8	A	overall	0.82	0.74
ELA	8	B	overall	0.78	0.69
ELA	8	C	overall	0.80	0.72
ELA	8	D	overall	0.79	0.71
ELA	8	E	overall	0.79	0.70
ELA	8	F	overall	0.79	0.71
ELA	8	G	overall	0.81	0.73
ELA	8	H	overall	0.80	0.73
ELA	8	A	1 vs 2,3,4	0.91	0.87
ELA	8	B	1 vs 2,3,4	0.94	0.91
ELA	8	C	1 vs 2,3,4	0.94	0.91
ELA	8	D	1 vs 2,3,4	0.93	0.90
ELA	8	E	1 vs 2,3,4	0.93	0.90
ELA	8	F	1 vs 2,3,4	0.94	0.92
ELA	8	G	1 vs 2,3,4	0.94	0.91
ELA	8	H	1 vs 2,3,4	0.93	0.91
ELA	8	A	1,2 vs 3,4	0.92	0.89
ELA	8	B	1,2 vs 3,4	0.88	0.83
ELA	8	C	1,2 vs 3,4	0.89	0.85
ELA	8	D	1,2 vs 3,4	0.88	0.84
ELA	8	E	1,2 vs 3,4	0.88	0.84
ELA	8	F	1,2 vs 3,4	0.88	0.83
ELA	8	G	1,2 vs 3,4	0.90	0.86
ELA	8	H	1,2 vs 3,4	0.89	0.85
ELA	8	A	1,2,3 vs 4	0.99	0.98
ELA	8	B	1,2,3 vs 4	0.97	0.95
ELA	8	C	1,2,3 vs 4	0.97	0.96
ELA	8	D	1,2,3 vs 4	0.97	0.97
ELA	8	E	1,2,3 vs 4	0.97	0.96
ELA	8	F	1,2,3 vs 4	0.96	0.95
ELA	8	G	1,2,3 vs 4	0.98	0.97
ELA	8	H	1,2,3 vs 4	0.98	0.97

Table 20.28: Decision Consistency and Accuracy for Grade 10 ELA

Subject	Grade	Form	Cut	Accuracy	Consistency
ELA	10	A	overall	0.79	0.71
ELA	10	B	overall	0.80	0.71
ELA	10	C	overall	0.78	0.70
ELA	10	D	overall	0.79	0.71
ELA	10	E	overall	0.78	0.69
ELA	10	F	overall	0.79	0.70
ELA	10	G	overall	0.80	0.73
ELA	10	H	overall	0.80	0.72
ELA	10	A	1 vs 2,3,4	0.90	0.86
ELA	10	B	1 vs 2,3,4	0.93	0.89
ELA	10	C	1 vs 2,3,4	0.93	0.90
ELA	10	D	1 vs 2,3,4	0.93	0.90
ELA	10	E	1 vs 2,3,4	0.93	0.89
ELA	10	F	1 vs 2,3,4	0.93	0.90
ELA	10	G	1 vs 2,3,4	0.92	0.89
ELA	10	H	1 vs 2,3,4	0.93	0.90
ELA	10	A	1,2 vs 3,4	0.91	0.88
ELA	10	B	1,2 vs 3,4	0.90	0.85
ELA	10	C	1,2 vs 3,4	0.88	0.83
ELA	10	D	1,2 vs 3,4	0.89	0.85
ELA	10	E	1,2 vs 3,4	0.88	0.83
ELA	10	F	1,2 vs 3,4	0.89	0.84
ELA	10	G	1,2 vs 3,4	0.91	0.87
ELA	10	H	1,2 vs 3,4	0.90	0.85
ELA	10	A	1,2,3 vs 4	0.98	0.97
ELA	10	B	1,2,3 vs 4	0.97	0.96
ELA	10	C	1,2,3 vs 4	0.97	0.96
ELA	10	D	1,2,3 vs 4	0.97	0.96
ELA	10	E	1,2,3 vs 4	0.97	0.96
ELA	10	F	1,2,3 vs 4	0.97	0.96
ELA	10	G	1,2,3 vs 4	0.98	0.97
ELA	10	H	1,2,3 vs 4	0.97	0.97

20.5.1 Interpretation Considerations

Several factors might affect decision consistency and accuracy. Lower consistency and accuracy occurs when there are more performance levels because there is more opportunity for misclassification. Another important factor is the reliability of the scores. All other things being equal, more reliable test scores result in more similar reclassifications and less measurement error. Another factor is the location of the cut scores in the score distribution. More consistent and accurate classifications are observed when the cut scores are located away from the mass of the score distribution. For example, if scores are normally distributed, the mass is concentrated in the middle of the distribution; and thus, classifications become more consistent as cut scores go up (e.g., from 70 to 80 percent correct, from 80 to 90 percent correct, etc.).

Or alternatively, go down from 30 to 20 percent correct, from 20 to 10 percent correct, etc.

20.5.2 Results and Observations

Across all subject areas, the overall decision consistency ranged from the mid-0.60s to the high 0.70s; the decision accuracy ranged from the mid 0.70s to the mid-0.80s. The overall consistency and accuracy in ELA was slightly lower than mathematics, on average. As expected, the consistency and accuracy indices for four performance levels are lower than the indices based on two categories. Dichotomous decisions using the Level 1/2 cuts generally have the highest consistency and accuracy values and exceeded 0.90 in all cases. The next highest values, on average, are associated with the Level 2/3 cut score for ELA and the Level 3/4 cut score in mathematics.

20.6 Rater Agreement

This year there were no constructed-response items that required human scorers. However, there will be in the future. When human scoring occurs, this will be another source of random error to evaluate. Test score reliability differs from scorer reliability and the need for one kind of estimate cannot be satisfied by the other.[6] The most easily obtainable data that captures rater consistency will come from *read behinds* collected during the scoring process. Consequently, future technical reports will likely document *rater agreement*.

[6] Frisbie (2005)

Other indicators, such as $inter-rater\,reliability$ in the form of interclass correlations, might be provided as well.

21
Operational Test Statistics

This chapter presents various summary statistics for KAP total test scores. As a critical quality-control activity, summary statistics should be monitored on an ongoing basis to detect unusual changes that might warrant further investigation.

Item characteristics affect total test-score characteristics. The chapters on classical and IRT item statistics provide information about item difficulty and item discrimination.

21.1 Demographic Information

The demographic profile of students in the state should be monitored along with the total test statistics because shifts in the state's demographics might affect test-score distributions. Demographic results for math and ELA are presented separately but are extremely similar, as expected. There was some fluctuation in the proportion of students across grades, but the difference was generally no more than 0.02 to 0.03 in value.

The proportion of White students in Grade 6 differed between math and ELA.

During this test administration, the largest racial group was White, which constituted about four-fifths of students. African American students made up about 10% of the population. Regarding ethnicity, about one-fifth of the students were Hispanic. ESOL (English speakers of other languages) students made up just over 10% of the population as did SWDs (students with disabilities).

Group by Grade	3	4	5	6	7	8	10
AmericanIndian	0.03	0.04	0.04	0.04	0.04	0.04	0.04
Asian	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Multi	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NativeHIorPacIsl	0.00	0.00	0.00	0.00	0.00	0.00	0.00
White	0.80	0.80	0.79	0.79	0.79	0.80	0.81
Hispanic	0.20	0.19	0.19	0.19	0.18	0.18	0.17
ESOL	0.13	0.13	0.13	0.12	0.12	0.11	0.09
SWD	0.12	0.12	0.12	0.12	0.12	0.11	0.10

Table 21.1: Proportion of Students in Demographic Groups by Grade for Math

Group by Grade	3	4	5	6	7	8	10
AfricanAmerican	0.08	0.07	0.07	0.07	0.07	0.07	0.07
AmericanIndian	0.03	0.04	0.04	0.04	0.04	0.04	0.04
Asian	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Multi	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NativeHIorPacIsl	0.00	0.00	0.00	0.00	0.00	0.00	0.00
White	0.80	0.80	0.79	0.80	0.79	0.80	0.81
Hispanic	0.19	0.19	0.19	0.19	0.18	0.18	0.17
ESOL	0.13	0.13	0.13	0.12	0.11	0.11	0.09
SWD	0.12	0.12	0.12	0.12	0.12	0.11	0.10

Table 21.2: Proportion of Students in Demographic Groups by Grade for ELA

21.2 Performance Level Statistics

The KAP classifies students into four performance levels: Level 1, Level 2, Level 3, and Level 4. For accountability purposes, the combined proportion of students in Level 3 and Level 4 is important. The proportions of students in each performance level, as well as the combined Level 3 and Level 4 proportion are provided here. The dominant feature in the performance trends across grades in both subjects was a general decline in the combined proportion of students in Level 3 and Level 4. This decline was present in math Grades 3–8 and in ELA Grades 4–8. Because this is the initial baseline year for KAP, one would expect these values to change in future administrations, ideally with increases in the combined Level 3 and Level 4 proportions.

Grade	Level 1	Level 2	Level 3	Level 4	Level 3 + 4
3	0.122	0.353	0.366	0.160	0.525
4	0.135	0.502	0.280	0.083	0.363
5	0.231	0.426	0.238	0.105	0.342
6	0.203	0.462	0.249	0.086	0.335
7	0.150	0.551	0.266	0.034	0.299
8	0.363	0.401	0.194	0.042	0.237
10	0.372	0.381	0.199	0.048	0.247

Table 21.3: Proportion of Students in Each Performance Level by Grade for Math

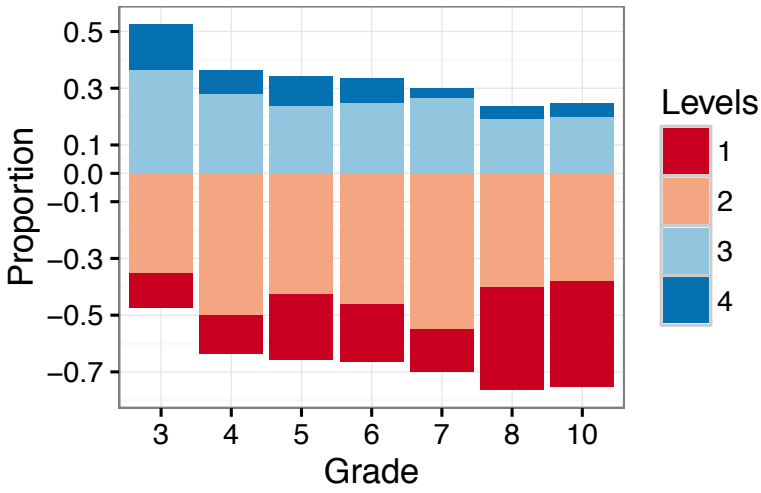


Figure 21.1: Peformance-Level Results in Math

Grade	Level 1	Level 2	Level 3	Level 4	Level 3 + 4
3	0.196	0.327	0.345	0.132	0.477
4	0.105	0.333	0.451	0.112	0.563
5	0.176	0.328	0.347	0.149	0.496
6	0.265	0.325	0.372	0.038	0.410
7	0.251	0.344	0.376	0.029	0.405
8	0.203	0.493	0.281	0.023	0.304
10	0.240	0.442	0.296	0.022	0.318

Table 21.4: Proportion of Students in Each Performance Level by Grade for ELA

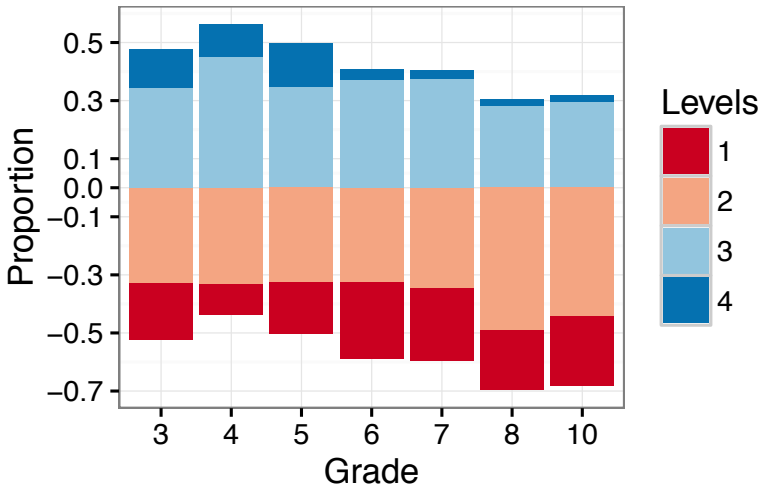


Figure 21.2: Peformance-Level Results in ELA

21.3 Scaled Scores

Summary statistics for total-test scaled scores are provided here. These results are presented together for simplicity, but they should not be compared. Longitudinal trends in the summary statistics are a different matter; future test results can, and should, be compared to prior test results. However, such comparisons must be limited to the same subject and grade (e.g., Grade 4 math results in 2016 can be compared to Grade 4 math results in 2015).

As emphasized in the scaling chapter, each test was scaled separately, so the results cannot be compared. The fact that one grade's mean scaled score is higher than another's is of no practical importance.

A few observations are noteworthy. All standard deviation values were near 25, which is expected given the KAP scaling procedures (outlined in the scaling chapter). The minimum and maximum values are within the LOSS (lowest observable scaled score) and HOSS (highest observable scaled score) values of 220 and 380, respectively. The shapes of the distributions can be ascertained by comparing the differences between (1) P_{50} and P_{25} and (2) P_{75} and P_{50}. The larger of the two differences will indicate the direction of any skew in the distribution (a negative skew when the first difference is larger and a positive skew when the second difference is larger). If the two differences are the same, the distribution is symmetric. In ELA, the distributions are more symmetric in shape. In math, the distributions at higher grade levels were positively skewed.

The maximum scaled score for some ELA tests did reach the HOSS value. This is permissible because in any given administration year, some tests may not reach the LOSS or HOSS value. This year's results meet the requirement that no test should have a minimum scaled score less than 220 or a maximum scaled score greater than 380.

Grade	Mean	SD	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
3	303.2	24.4	220	273	284	301	319	337	380
4	293.0	24.7	220	263	275	291	309	326	380
5	292.2	24.5	220	263	273	289	307	326	380
6	292.6	23.9	220	265	275	289	306	326	380
7	289.6	24.0	220	262	272	286	304	323	380
8	285.7	23.9	220	260	268	281	299	319	380
10	285.7	23.7	220	259	269	281	299	319	380

Table 21.5: Scaled-Score Descriptive Statistics by Grade for Math

Grade	Mean	SD	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
3	298.3	24.7	220	266	279	298	316	331	380
4	303.6	24.9	220	270	285	304	322	336	378
5	298.6	25.0	220	265	281	299	317	330	373
6	292.9	24.7	220	260	276	293	311	325	371
7	291.3	25.1	220	256	274	293	310	323	363
8	285.9	24.7	220	253	268	286	304	318	364
10	286.7	24.7	220	254	269	287	304	319	360

Table 21.6: Scaled-Score Descriptive Statistics by Grade for ELA

21.4 Longitudinal Trends

THE FORMAT OF THE previously presented results is not ideal for longitudinal monitoring. Future technical manuals will include an appendix that will document results in a manner that facilitates cross-year comparisons for quality-control purposes.

A longitudinal trend table might look like the following table. Each grade and subject-area test will have its own table. This year's test results will serve as the baseline. Key statistics will constitute the rows of the tables. The columns will include the results from each administration year. Tiny graphs, known as *sparklines*, will allow quick visual inspection of the trend for each statistic.

Table 21.7: Possible Format for a Future Data Trend Table

Statistic	2015	2016	2017	2018	Trend Line
Total N	37720	36961	37737	37196	
AfricanAmerican	0.0754	0.0704	0.0754	0.0724	
AmericanIndian	0.0303	0.0403	0.0305	0.0367	
Asian	0.0279	0.029	0.0283	0.028	
Multi	0.0009	0.001	0.0009	0.001	
NativeHIorPacIsl	0.0023	0.0028	0.0022	0.0023	
White	0.8018	0.7945	0.8012	0.7979	
Hispanic	0.195	0.1916	0.196	0.1931	
ESOL	0.1274	0.1261	0.1293	0.1274	
SWD	0.1227	0.122	0.1226	0.1246	
Mean	298.3	298.63	303.17	303.56	
SD	24.67	24.97	24.38	24.87	
Min	220	220	220	220	
P-10	266	265	273	270	
P-25	279	281	284	285	
P-50	298	299	301	304	
P-75	316	317	319	322	
P-90	331	330	337	336	
Max	380	373	380	378	
Level_1	0.1962	0.1762	0.122	0.1048	
Level_2	0.3266	0.3276	0.3528	0.3327	
Level_3	0.3452	0.347	0.3656	0.451	
Level_4	0.1321	0.1493	0.1596	0.1115	
Level_3_4	0.4773	0.4963	0.5252	0.5625	
SSCut1_2	276	276	276	276	
SSCut2_3	300	300	300	300	
SSCut3_4	327	327	327	327	
ThetaCut1_2	-1.015	-1.015	-1.015	-1.015	
ThetaCut2_3	-0.05	-0.05	-0.05	-0.05	
ThetaCut3_4	1.02	1.02	1.02	1.02	

Part VI

Technical Quality: Validity

22
Validity

As defined in the *Standards for Educational and Psychological Testing*, validity refers to: *the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests.*[1]

[1] AERA, APA, & NCME, 2014, p. 11

The word *validity* is sometimes used in isolation for brevity. As the Standards' definition makes clear, validity always concerns the uses and interpretations of test scores, even if they are not always explicitly stated.

The Standards provide a framework for describing the sources of evidence that should be considered when evaluating test-score validity. These sources include evidence based on 1) test content, 2) response processes, 3) internal test structure, 4) relationships between test scores and other variables, and 5) consequences of testing.[2]

[2] See AERA, APA, & NCME, 2014, p. 13—21

Other sources of evidence also can bolster the validity argument. For example, when item-response theory (IRT) is used to analyze assessment data, validity considerations related to the use of IRT should be explored. When cut scores are critical to the interpretation of test results, the procedural validity of the processes used to establish those scores also should be addressed.

The validation process involves the ongoing collection of a variety of evidence to support the proposed test-score interpretations and uses. Much of this technical manual describes aspects of the KAP tests that support KAP test score interpretations and uses. These include the chapters on item and test development, test administration, test scoring, standard setting, item analysis, IRT calibration, scaling, linking, score reporting, and reliability. The information that follows summarizes and synthesizes the validity evidence based on the Standards' framework. The purposes of the KAP tests, and the intended uses of the KAP test scores, are reviewed first, and then each type of validity evidence is addressed in turn.

Some of this technical manual's content is not directly part of the validity argument but provides information that will be important for peer review.

22.1 Purposes of KAP and Intended Uses of KAP Scores

The Standards emphasize that validity concerns how test scores are used. To contextualize the evidence that will be presented below, the

purposes of the KAP will be reviewed first. As stated in the technical manual's introduction, the purposes of the KAP include the following:

- Measure specific claims related to the *Kansas College and Career Ready Standards* (KCCRS) as identified in the Performance Level Descriptors (PLDs); * Provide information for calculating Annual Measurable Objectives (AMOs) and for state accreditation;
- Report individual student scores and performance levels; and
- Provide subscale and total scores that, when used with local assessment scores, can help improve building's or district's programs in the tested content areas.

22.2 Evidence Based on Test Content

Test content validity evidence for the KAP rests greatly on establishing a link between each piece of the assessment (e.g., the items) and what students should know and be able to do as described in the KCCRS. Thus, the evidence supporting the alignment among the KAP tasks and the KCCRS should be provided.

Multiple procedural steps can be used to evaluate the content validity of the KAP.

- Evaluate the degree to which the KAP test specifications represent and align with the knowledge and skills described in the KCCRS.
- Evaluate the alignment between the KAP items and test specifications to ensure representativeness.
- Evaluate the extent to which the curriculum aligns with the KCCRS.
- Conduct content reviews of the KAP items using a panel of content experts to see if the items measure the intended construct or if sources of construct-irrelevant variance exist.
- Conduct fairness reviews of KAP items to avoid bias and sensitivity issues related to specific subpopulations.
- Evaluate procedures for KAP test administration and scoring, such as the appropriateness of instructions to examinees, time limits for the assessment, etc.
- Submit operational tests for third-party, independent reviews.

For example, if some KAP test content is not included in the curriculum, then low scores on the KAP should not be interpreted to mean that instruction was ineffective.

Several chapters in the first half of this technical manual present validity evidence related to test content. As described in those chapters, all the KAP items were developed and aligned with the KCCRS, and item development followed well-established procedures. After the items were developed, they underwent multiple rounds of content and bias reviews. After testing, the items were reviewed with respect to

their statistical properties. Items selected as operational items had to pass content, psychometric, and KSDE reviews. Tests also were administered according to standardized procedures, with allowable accommodations.

Specific efforts to ensure content validity are summarized below.

- AAI used Webb's (1999) Depth of Knowledge (DOK) model to ensure that KAP items aligned with the KCCRS in terms of both content and cognitive levels.
- AAI ensured that (1) detailed test and item/passage development specifications were established, (2) tests included sufficient numbers of items, and (3) items were adequately distributed across content, levels of cognitive complexity, and difficulty.
- AAI selected qualified item writers and provided training to ensure they wrote high-quality items.
- Each newly developed item was first reviewed by content specialists and editors at AAI to make sure that all items were aligned with the KCCRS. Appropriateness for the intended grade was also considered, as well as depth of knowledge, graphics, grammar/punctuation, language demand, and distractor reasonableness.
- Test items were submitted for review to content committees composed of Kansas educators who considered, but were not limited to, the following categories:
  - Overall quality and clarity
  - KCCRS alignment
  - Grade-level appropriateness
  - Difficulty level
  - Depth of knowledge
  - Appropriate sources of challenge (e.g., item difficulty was not related to unintended content or skills)
  - Correct answer
  - Quality of distractors
  - Graphics
  - Appropriate language demand
  - Freedom from bias
- An external bias, fairness, and sensitivity committee reviewed items for issues related to diversity, gender, and other factors.
- Several statistical analyses were conducted before items were selected for operational use, including classical item analysis, distractor analysis, and differential item functioning (DIF). AAI staff again carefully reviewed items' statistical characteristics. DIF was used to detect items that might bias test scores for particular groups.

Empirical investigation of DIF strengthens the validity evidence related to score interpretations for students in particular groups by eliminating potential sources of construct-irrelevant variance; thus, DIF results might be better considered as internal–structure validity evidence.

- Administration of the KAP tests was standardized and included allowable testing accommodations. Students were given ample time to complete the tests (i.e., there were no speededness issues).

Once multidisciplinary performance tasks (MDPTs) become operational, their raters will be carefully recruited and well trained. Scoring will be monitored throughout the scoring window to ensure that an acceptable level of accuracy is maintained.

22.2.1 Alignment

Results of an independent alignment study for KAP should be available in the 2016 technical manual.

22.3 Evidence Based on Response Processes

Response-process evidence examines the extent to which the cognitive skills and processes employed by students match those identified in the construct domains defined by test developers for all students and for each subgroup. Think-aloud procedures or cognitive labs can be used to collect this type of evidence. Currently no such studies have been conducted for the KAP. If and when such studies occur, the results will be included in the technical manual.

When the MDPTs become operational, an examination of the extent to which the raters interpret and apply the scoring criteria accurately when assigning scores to students' responses would also provide response–process evidence for validity.

22.4 Evidence Based on Internal Structure

As described in the Standards (2014), internal-structure evidence refers to the degree to which relationships between test items and test components conform to the construct the intended test uses and on which interpretations are based. For each KAP test, one total test score and separate claim scores are reported (see the chapter on item and test scores for more information). Several sources of evidence provide internal-structure evidence relating to the use of both types of scores.

22.4.1 Item-Test Correlations

Item-test correlations (indicators of item discrimination) are reviewed in the chapter on classical item statistics. Most items had acceptable discrimination values. However, some extremely difficult items had low discrimination values that were likely attenuated by their difficulty.

When a test measures new standards, it is not unusual for some items to be very difficult. Item difficulty and discrimination will be monitored in the coming years to see how these item statistics change.

22.4.2 IRT Dimensionality

Results from the IRT dimensionality study were presented in the IRT chapter. The KAP tests were essentially unidimensional, providing evidence supporting interpretations based on the total scores for the respective KAP tests.

One might expect some dimensionality issues to occur when the MDPTs become operational next year.

Because the ELA tests were composed in part by testlets (items associated with a common stimulus such as a reading passage), the presence of some minor dimensions might have been expected. However, even these tests had a dominant first dimension. Further, although some concerns appeared with respect to local item dependence, those were no more significant in ELA than math.

22.4.3 Added Value of Subscores

The ELA and Mathematics tests administered in 2015 as part of the KAP included four subscale scores per subject, called claim scores. Claim scores can be of added value if they provide information about specific abilities or proficiency than the more general total score. It is important to evaluate the psychometric quality of claim scores because they do not always add value (new information) beyond that provided by the total score. Claim scores may be less reliable than total scores because claim scores are based on fewer items and may not be orthogonal enough to the total test to add unique information. One way to measure the added value of a subscale is using an equation from Feinberg and Wainer.[3]

In ELA, Claim 1 had enough items to report breakout scores over informational and literary texts. In this section, ELA Claims 2 and 3 represent those scores.

[3] Feinberg and Wainer, 2014

22.4.3.1 Feinberg and Wainer's Equation

Haberman[4] describes theory-driven, regression–based methods for evaluating the value of a subscore. However, his statistical evaluation criteria were derived from the assumption that subscores were computed in one of three specific ways. For other situations, Haberman's methods may not apply.

[4] Haberman, 2008

KAP subscores were not computed in any of the ways Haberman describes.

Feinberg and Wainer present an equation derived from Feinberg's (2012)[5] extensive simulations, rather than theory, but applies to the evaluation of subscores regardless of how they are computed. The equation evaluates subscores based on their reliability and orthogonality to the rest of the test. Thus, this strategy was selected for evaluating KAP claim scores.

[5] Cited in Feinberg and Wainer, 2014

Feinberg and Wainer's *value added ratio* (VAR) is equal to $1.15 + 0.51(r_1) + 0.67(r_2)$, where r_1 is the reliability of the subscore and r_2 is the raw (Pearson) correlation between the subscore and the remainder of the test, divided by the square root of the product of their reliabilities. The VAR is an approximation to Haberman's ratio of proportional reduction in mean–square error values for the subscore over the total score. Thus, VAR < 1 means the subscore is *less* valuable than the total score, and VAR > 1 *means* the subscore is more valuable than the total score.

Because KAP claim scores were computed using item response theory, we used the IRT-based marginal reliability for r_1 and expected

a posteriori (EAP) scores in the correlation for r_2. Otherwise, classical test theory (CTT) statistics were used. Specifically, summed scores were used for scores on the remainder of the test in r_2, and two coefficient alpha values constituted the reliability statistics for the denominator of r_2. A VAR value was computed for each of four Claims on eight forms (A–H) in Grades 3–8 and 10 for ELA and for Math.

22.4.3.2 Added Value of Claim Scores

The following tables list the VAR for each claim score for every form administered in every grade. The number of items (k) is given for each claim score. All VAR values for both ELA and Math are less than 1, indicating that none of the claim scores add value to the total score. It should be noted that the Feinberg and Wainer equation has been criticized[6] as being too simple; however, it may be the only available tool for the present situation because 1) the KAP claim scores were not computed using one of the methods described by Haberman and 2) an external criterion is not currently available to evaluate KAP claim scores against. In the future, additional information may be available for evaluating the quality of KAP claim scores, or claim scores may be computed in a different way.

[6] Sinharay, 2015

Table 22.1: Added Value Analysis for ELA Claims: Grade 3

Grade	Form	Claim	k	IRT r Claim	Alpha r Claim	Corr.	Alpha r Rest	r_2	VAR
3	A	2	19	0.78	0.80	0.84	0.89	0.99	0.89
3	B	2	19	0.78	0.79	0.80	0.85	0.98	0.90
3	C	2	17	0.74	0.76	0.79	0.86	0.98	0.87
3	D	2	18	0.78	0.78	0.79	0.84	0.97	0.90
3	E	2	19	0.79	0.79	0.81	0.88	0.97	0.90
3	F	2	18	0.73	0.73	0.77	0.86	0.97	0.87
3	G	2	27	0.82	0.80	0.82	0.86	0.99	0.90
3	H	2	18	0.73	0.73	0.78	0.88	0.97	0.87
3	A	3	24	0.73	0.77	0.81	0.90	0.97	0.87
3	B	3	18	0.75	0.76	0.78	0.87	0.96	0.89
3	C	3	24	0.74	0.75	0.79	0.86	0.98	0.87
3	D	3	17	0.70	0.73	0.74	0.86	0.94	0.88
3	E	3	26	0.75	0.77	0.80	0.88	0.97	0.88
3	F	3	26	0.78	0.79	0.78	0.84	0.95	0.91
3	G	3	18	0.75	0.78	0.79	0.88	0.95	0.89
3	H	3	26	0.78	0.78	0.78	0.86	0.95	0.91
3	A	4	25	0.80	0.84	0.84	0.88	0.98	0.90
3	B	4	18	0.74	0.74	0.78	0.87	0.97	0.88
3	C	4	24	0.77	0.78	0.80	0.86	0.98	0.89
3	D	4	17	0.72	0.75	0.76	0.86	0.95	0.88
3	E	4	23	0.80	0.80	0.82	0.88	0.98	0.90
3	F	4	24	0.78	0.75	0.80	0.86	1.00	0.88
3	G	4	22	0.75	0.76	0.80	0.88	0.98	0.87
3	H	4	24	0.77	0.79	0.79	0.86	0.96	0.90

Table 22.2: Added Value Analysis for ELA Claims: Grade 4

Grade	Form	Claim	k	IRT r Claim	Alpha r Claim	Corr.	Alpha r Rest	r_2	VAR
4	A	2	19	0.74	0.80	0.83	0.91	0.97	0.88
4	B	2	19	0.73	0.77	0.77	0.83	0.96	0.88
4	C	2	26	0.80	0.83	0.80	0.82	0.97	0.91
4	D	2	18	0.71	0.74	0.74	0.80	0.97	0.86
4	E	2	27	0.80	0.82	0.80	0.84	0.97	0.91
4	F	2	19	0.74	0.76	0.79	0.88	0.96	0.88
4	G	2	28	0.80	0.82	0.80	0.83	0.97	0.91
4	H	2	19	0.75	0.76	0.78	0.86	0.96	0.88
4	A	3	27	0.80	0.84	0.83	0.90	0.96	0.92
4	B	3	18	0.69	0.70	0.74	0.85	0.95	0.86
4	C	3	18	0.67	0.69	0.76	0.88	0.98	0.83
4	D	3	18	0.64	0.66	0.71	0.83	0.96	0.83
4	E	3	18	0.66	0.69	0.76	0.88	0.97	0.84
4	F	3	25	0.75	0.78	0.80	0.87	0.97	0.89
4	G	3	17	0.68	0.70	0.75	0.87	0.96	0.86
4	H	3	24	0.76	0.77	0.79	0.85	0.98	0.88
4	A	4	22	0.79	0.85	0.84	0.90	0.95	0.91
4	B	4	15	0.72	0.73	0.76	0.84	0.97	0.87
4	C	4	20	0.71	0.71	0.77	0.88	0.97	0.86
4	D	4	17	0.69	0.69	0.74	0.82	0.98	0.84
4	E	4	22	0.74	0.76	0.78	0.86	0.96	0.89
4	F	4	23	0.77	0.79	0.80	0.86	0.97	0.89
4	G	4	19	0.71	0.72	0.77	0.87	0.97	0.86
4	H	4	19	0.73	0.73	0.77	0.86	0.97	0.87

Table 22.3: Added Value Analysis for ELA Claims: Grade 5

Grade	Form	Claim	k	IRT r Claim	Alpha r Claim	Corr.	Alpha r Rest	r_2	VAR
5	A	2	18	0.69	0.79	0.83	0.89	0.98	0.85
5	B	2	18	0.71	0.74	0.75	0.81	0.97	0.86
5	C	2	26	0.79	0.81	0.79	0.82	0.96	0.91
5	D	2	17	0.65	0.68	0.70	0.80	0.94	0.85
5	E	2	26	0.76	0.78	0.79	0.84	0.97	0.89
5	F	2	25	0.77	0.76	0.78	0.82	0.98	0.89
5	G	2	17	0.72	0.75	0.78	0.85	0.97	0.87
5	H	2	26	0.77	0.78	0.80	0.83	0.98	0.88
5	A	3	26	0.69	0.79	0.83	0.89	0.99	0.84
5	B	3	18	0.63	0.68	0.70	0.84	0.93	0.85
5	C	3	13	0.57	0.60	0.69	0.88	0.95	0.80
5	D	3	17	0.62	0.70	0.66	0.80	0.89	0.87
5	E	3	17	0.63	0.64	0.72	0.88	0.97	0.82
5	F	3	17	0.68	0.69	0.74	0.85	0.96	0.85
5	G	3	26	0.71	0.74	0.77	0.86	0.97	0.86
5	H	3	17	0.68	0.70	0.75	0.87	0.97	0.85
5	A	4	23	0.75	0.82	0.83	0.88	0.98	0.88
5	B	4	16	0.69	0.72	0.73	0.83	0.94	0.87
5	C	4	22	0.75	0.77	0.79	0.85	0.97	0.88
5	D	4	16	0.65	0.66	0.68	0.81	0.92	0.86
5	E	4	21	0.77	0.79	0.80	0.84	0.98	0.89
5	F	4	20	0.71	0.71	0.75	0.84	0.97	0.86
5	G	4	22	0.75	0.76	0.77	0.85	0.95	0.89
5	H	4	20	0.72	0.74	0.77	0.86	0.96	0.87

Table 22.4: Added Value Analysis for ELA Claims: Grade 6

Grade	Form	Claim	k	IRT r Claim	Alpha r Claim	Corr.	Alpha r Rest	r_2	VAR
6	A	2	17	0.74	0.81	0.84	0.90	0.98	0.87
6	B	2	17	0.69	0.70	0.73	0.82	0.97	0.85
6	C	2	26	0.80	0.81	0.80	0.82	0.98	0.90
6	D	2	17	0.71	0.75	0.75	0.82	0.95	0.87
6	E	2	16	0.73	0.73	0.76	0.86	0.96	0.88
6	F	2	18	0.74	0.76	0.77	0.85	0.96	0.88
6	G	2	16	0.76	0.77	0.79	0.85	0.97	0.89
6	H	2	23	0.79	0.80	0.79	0.81	0.98	0.90
6	A	3	26	0.77	0.81	0.83	0.90	0.98	0.89
6	B	3	16	0.68	0.68	0.74	0.82	0.99	0.83
6	C	3	15	0.69	0.65	0.75	0.87	0.99	0.83
6	D	3	16	0.72	0.74	0.74	0.83	0.94	0.89
6	E	3	22	0.74	0.76	0.78	0.85	0.98	0.87
6	F	3	24	0.74	0.75	0.76	0.85	0.95	0.89
6	G	3	25	0.78	0.79	0.80	0.84	0.98	0.89
6	H	3	16	0.69	0.71	0.76	0.85	0.97	0.85
6	A	4	25	0.76	0.82	0.84	0.89	0.98	0.88
6	B	4	18	0.70	0.72	0.74	0.81	0.97	0.86
6	C	4	21	0.73	0.74	0.76	0.85	0.96	0.88
6	D	4	17	0.64	0.66	0.73	0.85	0.97	0.83
6	E	4	24	0.75	0.76	0.77	0.85	0.96	0.89
6	F	4	20	0.72	0.73	0.77	0.86	0.97	0.87
6	G	4	19	0.68	0.69	0.76	0.88	0.97	0.84
6	H	4	21	0.69	0.68	0.75	0.87	0.98	0.85

Table 22.5: Added Value Analysis for ELA Claims: Grade 7

Grade	Form	Claim	k	IRT r Claim	Alpha r Claim	Corr.	Alpha r Rest	r_2	VAR
7	A	2	18	0.76	0.78	0.85	0.90	1.01	0.86
7	B	2	18	0.73	0.70	0.78	0.81	1.03	0.83
7	C	2	26	0.78	0.79	0.81	0.86	0.97	0.89
7	D	2	17	0.68	0.71	0.70	0.77	0.95	0.86
7	E	2	17	0.69	0.73	0.77	0.87	0.96	0.86
7	F	2	24	0.72	0.74	0.78	0.83	0.98	0.86
7	G	2	16	0.70	0.73	0.77	0.86	0.97	0.86
7	H	2	24	0.72	0.76	0.77	0.83	0.97	0.87
7	A	3	26	0.78	0.84	0.85	0.88	0.99	0.89
7	B	3	15	0.66	0.67	0.74	0.82	1.00	0.82
7	C	3	18	0.71	0.73	0.78	0.88	0.96	0.86
7	D	3	13	0.55	0.56	0.66	0.82	0.97	0.78
7	E	3	25	0.75	0.78	0.79	0.85	0.97	0.88
7	F	3	17	0.64	0.65	0.73	0.86	0.97	0.82
7	G	3	23	0.73	0.74	0.77	0.86	0.97	0.87
7	H	3	15	0.60	0.61	0.70	0.87	0.96	0.81
7	A	4	21	0.72	0.81	0.84	0.90	0.98	0.86
7	B	4	17	0.68	0.71	0.73	0.81	0.96	0.85
7	C	4	23	0.76	0.79	0.80	0.87	0.97	0.89
7	D	4	16	0.65	0.68	0.68	0.79	0.93	0.85
7	E	4	21	0.75	0.77	0.78	0.86	0.96	0.89
7	F	4	23	0.75	0.76	0.75	0.83	0.95	0.90
7	G	4	24	0.74	0.77	0.78	0.85	0.96	0.88
7	H	4	24	0.76	0.79	0.77	0.82	0.95	0.90

Table 22.6: Added Value Analysis for ELA Claims: Grade 8

Grade	Form	Claim	k	IRT r Claim	Alpha r Claim	Corr.	Alpha r Rest	r_2	VAR
8	A	2	16	0.73	0.78	0.83	0.91	0.98	0.86
8	B	2	15	0.72	0.71	0.74	0.83	0.96	0.87
8	C	2	17	0.70	0.71	0.77	0.86	0.98	0.85
8	D	2	16	0.73	0.74	0.77	0.82	0.99	0.86
8	E	2	22	0.73	0.74	0.75	0.83	0.96	0.88
8	F	2	17	0.71	0.66	0.76	0.85	1.01	0.84
8	G	2	14	0.67	0.70	0.76	0.88	0.96	0.85
8	H	2	15	0.69	0.70	0.74	0.85	0.96	0.86
8	A	3	25	0.80	0.85	0.86	0.89	0.98	0.90
8	B	3	17	0.73	0.73	0.76	0.82	0.97	0.87
8	C	3	22	0.75	0.76	0.79	0.85	0.99	0.87
8	D	3	16	0.69	0.72	0.75	0.83	0.97	0.85
8	E	3	12	0.64	0.67	0.73	0.85	0.96	0.83
8	F	3	25	0.77	0.78	0.75	0.82	0.94	0.91
8	G	3	23	0.78	0.79	0.80	0.86	0.97	0.90
8	H	3	21	0.71	0.72	0.76	0.84	0.98	0.86
8	A	4	25	0.76	0.83	0.84	0.90	0.97	0.89
8	B	4	18	0.68	0.69	0.72	0.84	0.95	0.86
8	C	4	23	0.77	0.77	0.78	0.85	0.96	0.90
8	D	4	15	0.68	0.69	0.74	0.84	0.97	0.84
8	E	4	20	0.73	0.75	0.74	0.83	0.95	0.89
8	F	4	23	0.75	0.73	0.77	0.83	0.99	0.87
8	G	4	24	0.79	0.79	0.78	0.86	0.95	0.92
8	H	4	24	0.76	0.76	0.77	0.83	0.97	0.88

Table 22.7: Added Value Analysis for ELA Claims: Grade 10

Grade	Form	Claim	k	IRT r Claim	Alpha r Claim	Corr.	Alpha r Rest	r_2	VAR
10	A	2	18	0.75	0.77	0.81	0.89	0.98	0.88
10	B	2	26	0.81	0.81	0.81	0.85	0.98	0.91
10	C	2	17	0.74	0.75	0.76	0.86	0.94	0.89
10	D	2	17	0.72	0.73	0.79	0.89	0.97	0.86
10	E	2	16	0.72	0.74	0.77	0.85	0.97	0.87
10	F	2	25	0.78	0.79	0.78	0.84	0.95	0.91
10	G	2	24	0.76	0.76	0.79	0.87	0.97	0.88
10	H	2	17	0.75	0.75	0.78	0.86	0.98	0.88
10	A	3	22	0.76	0.81	0.82	0.88	0.98	0.88
10	B	3	18	0.74	0.76	0.79	0.87	0.97	0.88
10	C	3	26	0.76	0.80	0.76	0.84	0.93	0.92
10	D	3	26	0.78	0.80	0.79	0.87	0.94	0.92
10	E	3	21	0.72	0.75	0.77	0.84	0.96	0.87
10	F	3	17	0.64	0.68	0.74	0.87	0.96	0.83
10	G	3	17	0.76	0.78	0.78	0.87	0.95	0.90
10	H	3	25	0.79	0.81	0.78	0.84	0.95	0.92
10	A	4	24	0.74	0.80	0.79	0.88	0.94	0.90
10	B	4	21	0.70	0.74	0.76	0.88	0.95	0.88
10	C	4	18	0.69	0.72	0.74	0.87	0.93	0.88
10	D	4	24	0.78	0.82	0.79	0.87	0.93	0.92
10	E	4	22	0.74	0.75	0.77	0.85	0.97	0.88
10	F	4	21	0.77	0.77	0.78	0.85	0.97	0.90
10	G	4	22	0.78	0.79	0.80	0.87	0.97	0.90
10	H	4	22	0.75	0.73	0.78	0.88	0.98	0.88

Table 22.8: Added Value Analysis for Math Claims: Grade 3

Grade	Form	Claim	k	IRT r Claim	Alpha r Claim	Corr.	Alpha r Rest	r_2	VAR
3	A	1	44	0.88	0.87	0.77	0.73	0.97	0.95
3	B	1	36	0.87	0.84	0.75	0.68	0.99	0.93
3	C	1	44	0.90	0.89	0.77	0.69	0.98	0.95
3	D	1	44	0.90	0.89	0.81	0.77	0.98	0.95
3	E	1	35	0.88	0.86	0.75	0.69	0.98	0.94
3	F	1	44	0.90	0.89	0.77	0.69	0.99	0.94
3	G	1	44	0.89	0.88	0.79	0.75	0.97	0.96
3	H	1	44	0.90	0.90	0.82	0.75	0.99	0.94
3	A	2	8	0.56	0.49	0.68	0.89	1.03	0.75
3	B	2	5	0.39	0.37	0.59	0.87	1.03	0.66
3	C	2	8	0.46	0.49	0.64	0.90	0.97	0.73
3	D	2	7	0.52	0.50	0.68	0.91	1.01	0.74
3	E	2	6	0.49	0.50	0.65	0.88	0.98	0.75
3	F	2	6	0.42	0.44	0.62	0.90	0.98	0.71
3	G	2	7	0.42	0.45	0.60	0.90	0.95	0.73
3	H	2	8	0.57	0.54	0.71	0.91	1.01	0.76
3	A	3	6	0.51	0.49	0.64	0.89	0.98	0.76
3	B	3	6	0.49	0.44	0.65	0.87	1.05	0.69
3	C	3	4	0.45	0.35	0.64	0.91	1.13	0.63
3	D	3	7	0.57	0.58	0.73	0.91	1.00	0.77
3	E	3	4	0.44	0.43	0.65	0.88	1.06	0.67
3	F	3	5	0.52	0.48	0.68	0.90	1.04	0.72
3	G	3	6	0.54	0.54	0.69	0.90	0.99	0.76
3	H	3	5	0.47	0.47	0.67	0.91	1.02	0.70
3	A	4	6	0.42	0.43	0.59	0.89	0.95	0.73
3	B	4	4	0.40	0.39	0.58	0.87	1.00	0.69
3	C	4	7	0.40	0.41	0.60	0.91	0.97	0.70
3	D	4	7	0.52	0.50	0.68	0.91	1.01	0.74
3	E	4	6	0.44	0.37	0.61	0.89	1.06	0.66
3	F	4	6	0.47	0.37	0.65	0.91	1.13	0.63
3	G	4	6	0.46	0.48	0.64	0.90	0.97	0.74
3	H	4	6	0.48	0.47	0.67	0.91	1.02	0.71

Table 22.9: Added Value Analysis for Math Claims: Grade 4

Grade	Form	Claim	k	IRT r Claim	Alpha r Claim	Corr.	Alpha r Rest	r_2	VAR
4	A	1	45	0.90	0.90	0.80	0.73	0.99	0.95
4	B	1	34	0.87	0.86	0.78	0.67	1.03	0.90
4	C	1	45	0.90	0.88	0.78	0.69	1.00	0.94
4	D	1	45	0.90	0.89	0.82	0.75	1.01	0.94
4	E	1	36	0.89	0.87	0.75	0.69	0.97	0.96
4	F	1	43	0.91	0.89	0.80	0.75	0.98	0.96
4	G	1	45	0.90	0.89	0.83	0.76	1.01	0.93
4	H	1	44	0.90	0.88	0.83	0.78	1.00	0.94
4	A	2	7	0.50	0.46	0.66	0.92	1.02	0.72
4	B	2	3	0.19	0.22	0.42	0.89	0.95	0.61
4	C	2	7	0.42	0.42	0.59	0.90	0.96	0.72
4	D	2	7	0.51	0.41	0.67	0.91	1.10	0.67
4	E	2	6	0.38	0.44	0.57	0.89	0.91	0.73
4	F	2	7	0.50	0.46	0.67	0.91	1.04	0.71
4	G	2	7	0.49	0.42	0.67	0.91	1.09	0.67
4	H	2	6	0.48	0.42	0.66	0.91	1.06	0.68
4	A	3	7	0.46	0.42	0.62	0.92	1.01	0.71
4	B	3	4	0.47	0.44	0.66	0.88	1.07	0.68
4	C	3	7	0.49	0.48	0.67	0.89	1.02	0.71
4	D	3	8	0.65	0.60	0.75	0.91	1.01	0.80
4	E	3	4	0.56	0.46	0.71	0.88	1.11	0.69
4	F	3	7	0.53	0.49	0.67	0.91	1.01	0.74
4	G	3	8	0.65	0.60	0.77	0.91	1.04	0.78
4	H	3	7	0.58	0.55	0.73	0.90	1.04	0.75
4	A	4	8	0.51	0.52	0.67	0.91	0.97	0.76
4	B	4	5	0.52	0.49	0.65	0.88	0.98	0.76
4	C	4	7	0.41	0.40	0.60	0.90	1.01	0.68
4	D	4	5	0.45	0.42	0.63	0.91	1.02	0.70
4	E	4	5	0.36	0.39	0.50	0.89	0.86	0.76
4	F	4	8	0.53	0.52	0.69	0.91	1.00	0.75
4	G	4	5	0.45	0.44	0.62	0.91	0.98	0.73
4	H	4	7	0.60	0.60	0.73	0.90	0.98	0.80

Table 22.10: Added Value Analysis for Math Claims: Grade 5

Grade	Form	Claim	k	IRT r Claim	Alpha r Claim	Corr.	Alpha r Rest	r_2	VAR
5	A	1	45	0.90	0.87	0.80	0.76	0.99	0.95
5	B	1	36	0.88	0.87	0.77	0.71	0.98	0.94
5	C	1	44	0.90	0.89	0.80	0.75	0.99	0.95
5	D	1	44	0.91	0.90	0.75	0.68	0.96	0.97
5	E	1	36	0.89	0.87	0.70	0.61	0.96	0.96
5	F	1	44	0.90	0.87	0.74	0.69	0.96	0.97
5	G	1	44	0.91	0.90	0.78	0.74	0.96	0.97
5	H	1	44	0.90	0.90	0.81	0.75	0.98	0.95
5	A	2	7	0.56	0.52	0.69	0.90	1.01	0.76
5	B	2	5	0.53	0.51	0.68	0.89	1.01	0.74
5	C	2	6	0.43	0.43	0.62	0.91	0.99	0.71
5	D	2	6	0.33	0.34	0.56	0.91	1.01	0.64
5	E	2	4	0.29	0.27	0.53	0.89	1.08	0.57
5	F	2	5	0.34	0.34	0.52	0.89	0.95	0.69
5	G	2	7	0.44	0.43	0.64	0.92	1.01	0.70
5	H	2	8	0.54	0.50	0.70	0.91	1.04	0.73
5	A	3	8	0.62	0.59	0.71	0.90	0.98	0.82
5	B	3	5	0.40	0.41	0.61	0.89	1.02	0.67
5	C	3	8	0.60	0.60	0.72	0.90	0.98	0.80
5	D	3	7	0.50	0.46	0.67	0.91	1.03	0.72
5	E	3	5	0.43	0.41	0.64	0.88	1.07	0.65
5	F	3	6	0.53	0.53	0.69	0.89	1.01	0.74
5	G	3	7	0.50	0.51	0.65	0.92	0.95	0.77
5	H	3	7	0.48	0.50	0.68	0.91	1.01	0.72
5	A	4	8	0.51	0.43	0.65	0.90	1.03	0.72
5	B	4	6	0.41	0.38	0.61	0.89	1.04	0.66
5	C	4	7	0.52	0.46	0.67	0.91	1.04	0.72
5	D	4	7	0.45	0.43	0.63	0.91	1.00	0.71
5	E	4	5	0.35	0.32	0.55	0.88	1.02	0.64
5	F	4	7	0.42	0.43	0.59	0.89	0.96	0.72
5	G	4	6	0.48	0.51	0.66	0.91	0.98	0.74
5	H	4	6	0.49	0.47	0.68	0.91	1.04	0.71

Table 22.11: Added Value Analysis for Math Claims: Grade 6

Grade	Form	Claim	k	IRT r Claim	Alpha r Claim	Corr.	Alpha r Rest	r_2	VAR
6	A	1	46	0.89	0.88	0.67	0.56	0.95	0.97
6	B	1	34	0.85	0.82	0.70	0.60	0.99	0.92
6	C	1	43	0.89	0.88	0.72	0.64	0.97	0.96
6	D	1	46	0.90	0.88	0.75	0.66	0.99	0.95
6	E	1	37	0.88	0.86	0.69	0.55	1.00	0.93
6	F	1	41	0.87	0.86	0.69	0.54	1.02	0.91
6	G	1	44	0.90	0.90	0.77	0.67	1.00	0.94
6	H	1	37	0.87	0.85	0.71	0.59	1.01	0.92
6	A	2	7	0.28	0.30	0.45	0.89	0.87	0.71
6	B	2	6	0.45	0.44	0.61	0.84	1.01	0.70
6	C	2	7	0.38	0.39	0.57	0.89	0.96	0.70
6	D	2	6	0.42	0.40	0.62	0.89	1.04	0.66
6	E	2	4	0.31	0.25	0.48	0.88	1.02	0.62
6	F	2	6	0.38	0.35	0.55	0.87	1.00	0.67
6	G	2	6	0.48	0.43	0.68	0.90	1.09	0.67
6	H	2	5	0.39	0.34	0.54	0.86	1.00	0.68
6	A	3	8	0.36	0.28	0.52	0.89	1.04	0.64
6	B	3	5	0.32	0.30	0.51	0.85	1.00	0.64
6	C	3	7	0.37	0.37	0.57	0.89	0.99	0.67
6	D	3	6	0.37	0.36	0.57	0.89	1.02	0.66
6	E	3	5	0.38	0.32	0.58	0.87	1.10	0.61
6	F	3	6	0.25	0.22	0.45	0.87	1.02	0.59
6	G	3	8	0.45	0.41	0.64	0.90	1.06	0.67
6	H	3	6	0.42	0.37	0.61	0.86	1.09	0.64
6	A	4	8	0.32	0.26	0.47	0.88	0.97	0.66
6	B	4	5	0.23	0.22	0.40	0.85	0.94	0.64
6	C	4	7	0.31	0.32	0.52	0.89	0.98	0.65
6	D	4	6	0.40	0.39	0.58	0.89	0.98	0.70
6	E	4	3	0.28	0.25	0.47	0.87	1.01	0.61
6	F	4	5	0.30	0.25	0.49	0.87	1.04	0.61
6	G	4	7	0.37	0.28	0.57	0.91	1.13	0.59
6	H	4	5	0.29	0.26	0.49	0.87	1.05	0.60

Table 22.12: Added Value Analysis for Math Claims: Grade 7

Grade	Form	Claim	k	IRT r Claim	Alpha r Claim	Corr.	Alpha r Rest	r_2	VAR
7	A	1	42	0.87	0.87	0.76	0.67	1.00	0.92
7	B	1	35	0.85	0.85	0.68	0.55	1.00	0.91
7	C	1	44	0.84	0.83	0.76	0.70	1.00	0.91
7	D	1	44	0.87	0.85	0.75	0.66	1.00	0.92
7	E	1	35	0.83	0.83	0.75	0.67	1.01	0.90
7	F	1	45	0.87	0.87	0.76	0.69	0.98	0.93
7	G	1	42	0.86	0.85	0.77	0.72	0.99	0.93
7	H	1	32	0.84	0.83	0.78	0.73	1.01	0.91
7	A	2	7	0.49	0.46	0.62	0.88	0.97	0.74
7	B	2	5	0.36	0.41	0.56	0.86	0.94	0.70
7	C	2	8	0.47	0.56	0.64	0.86	0.92	0.77
7	D	2	8	0.53	0.51	0.67	0.87	1.01	0.74
7	E	2	6	0.53	0.49	0.68	0.85	1.05	0.72
7	F	2	8	0.51	0.58	0.67	0.88	0.94	0.78
7	G	2	7	0.52	0.57	0.69	0.88	0.98	0.76
7	H	2	6	0.53	0.58	0.68	0.87	0.96	0.78
7	A	3	7	0.41	0.35	0.59	0.89	1.07	0.64
7	B	3	5	0.33	0.22	0.54	0.87	1.22	0.50
7	C	3	7	0.47	0.41	0.65	0.87	1.09	0.66
7	D	3	6	0.38	0.30	0.57	0.88	1.11	0.60
7	E	3	5	0.33	0.25	0.55	0.87	1.19	0.53
7	F	3	6	0.43	0.38	0.62	0.89	1.06	0.66
7	G	3	6	0.42	0.40	0.60	0.88	1.01	0.69
7	H	3	5	0.42	0.35	0.60	0.88	1.07	0.65
7	A	4	4	0.49	0.39	0.67	0.89	1.14	0.64
7	B	4	2	0.40	0.36	0.59	0.86	1.06	0.64
7	C	4	4	0.51	0.40	0.66	0.87	1.12	0.65
7	D	4	4	0.46	0.29	0.63	0.87	1.25	0.55
7	E	4	5	0.53	0.45	0.68	0.85	1.09	0.69
7	F	4	6	0.48	0.45	0.65	0.88	1.03	0.70
7	G	4	6	0.46	0.40	0.64	0.88	1.07	0.67
7	H	4	5	0.54	0.46	0.68	0.87	1.08	0.70

Table 22.13: Added Value Analysis for Math Claims: Grade 8

Grade	Form	Claim	k	IRT r Claim	Alpha r Claim	Corr.	Alpha r Rest	r_2	VAR
8	A	1	46	0.88	0.88	0.67	0.53	0.98	0.94
8	B	1	35	0.85	0.83	0.66	0.56	0.97	0.93
8	C	1	40	0.86	0.86	0.69	0.60	0.96	0.95
8	D	1	45	0.87	0.85	0.71	0.63	0.98	0.94
8	E	1	35	0.84	0.83	0.63	0.48	1.00	0.91
8	F	1	44	0.88	0.87	0.68	0.52	1.01	0.92
8	G	1	43	0.87	0.86	0.72	0.61	1.00	0.93
8	H	1	35	0.87	0.84	0.68	0.54	1.01	0.92
8	A	2	6	0.20	0.31	0.33	0.89	0.63	0.83
8	B	2	5	0.15	0.31	0.37	0.86	0.71	0.75
8	C	2	4	0.20	0.32	0.45	0.88	0.86	0.68
8	D	2	4	0.28	0.35	0.53	0.87	0.96	0.65
8	E	2	4	0.20	0.18	0.42	0.85	1.10	0.51
8	F	2	6	0.21	0.24	0.44	0.88	0.97	0.61
8	G	2	4	0.26	0.34	0.53	0.88	0.98	0.63
8	H	2	3	0.15	0.26	0.38	0.86	0.79	0.70
8	A	3	5	0.27	0.24	0.48	0.89	1.05	0.58
8	B	3	4	0.25	0.20	0.45	0.85	1.08	0.55
8	C	3	6	0.33	0.31	0.53	0.88	1.02	0.63
8	D	3	5	0.35	0.31	0.53	0.87	1.02	0.64
8	E	3	5	0.30	0.22	0.51	0.84	1.17	0.51
8	F	3	5	0.31	0.25	0.52	0.88	1.11	0.56
8	G	3	7	0.34	0.29	0.54	0.88	1.07	0.60
8	H	3	5	0.30	0.34	0.52	0.86	0.97	0.65
8	A	4	6	0.41	0.34	0.57	0.88	1.04	0.66
8	B	4	5	0.46	0.43	0.58	0.85	0.96	0.75
8	C	4	6	0.45	0.42	0.57	0.87	0.95	0.74
8	D	4	6	0.50	0.44	0.60	0.87	0.97	0.76
8	E	4	3	0.37	0.24	0.53	0.84	1.17	0.55
8	F	4	4	0.42	0.30	0.60	0.88	1.17	0.58
8	G	4	6	0.48	0.42	0.63	0.87	1.03	0.70
8	H	4	4	0.39	0.30	0.54	0.86	1.06	0.64

Table 22.14: Added Value Analysis for Math Claims: Grade 10

Grade	Form	Claim	k	IRT r Claim	Alpha r Claim	Corr.	Alpha r Rest	r_2	VAR
10	A	1	42	0.87	0.87	0.65	0.51	0.97	0.94
10	B	1	37	0.87	0.87	0.70	0.62	0.96	0.95
10	C	1	41	0.87	0.87	0.70	0.61	0.96	0.95
10	D	1	43	0.86	0.86	0.64	0.51	0.97	0.94
10	E	1	42	0.88	0.88	0.69	0.55	0.99	0.93
10	F	1	40	0.87	0.86	0.64	0.56	0.93	0.97
10	G	1	45	0.87	0.88	0.66	0.54	0.96	0.95
10	H	1	35	0.87	0.87	0.66	0.56	0.95	0.96
10	A	2	5	0.32	0.33	0.53	0.87	0.98	0.65
10	B	2	8	0.45	0.46	0.62	0.87	0.97	0.73
10	C	2	8	0.42	0.45	0.58	0.88	0.93	0.74
10	D	2	6	0.34	0.36	0.54	0.86	0.96	0.68
10	E	2	5	0.33	0.31	0.54	0.88	1.04	0.62
10	F	2	6	0.33	0.36	0.53	0.87	0.95	0.68
10	G	2	5	0.32	0.29	0.52	0.88	1.04	0.62
10	H	2	7	0.39	0.41	0.56	0.87	0.95	0.71
10	A	3	5	0.30	0.29	0.52	0.88	1.03	0.62
10	B	3	5	0.26	0.25	0.51	0.88	1.09	0.55
10	C	3	4	0.38	0.37	0.57	0.88	1.00	0.67
10	D	3	5	0.30	0.22	0.54	0.87	1.24	0.48
10	E	3	4	0.36	0.40	0.61	0.88	1.02	0.65
10	F	3	5	0.30	0.29	0.50	0.87	1.01	0.63
10	G	3	5	0.27	0.29	0.52	0.88	1.03	0.60
10	H	3	4	0.33	0.28	0.54	0.87	1.08	0.60
10	A	4	4	0.16	0.20	0.37	0.88	0.89	0.63
10	B	4	5	0.22	0.26	0.45	0.88	0.95	0.62
10	C	4	6	0.21	0.26	0.42	0.88	0.88	0.67
10	D	4	3	0.13	0.16	0.36	0.87	0.96	0.57
10	E	4	6	0.17	0.17	0.43	0.89	1.12	0.49
10	F	4	5	0.18	0.22	0.40	0.87	0.91	0.63
10	G	4	5	0.20	0.24	0.44	0.88	0.96	0.61
10	H	4	6	0.18	0.17	0.40	0.88	1.04	0.54

22.4.3.3 Disattenuated Correlations

The values in the r_2 column of the prior tables are disattenuated correlations. Observed-score correlations are weakened by existing measurement error contained within each claim. As a result, disattenuating these correlations can provide an estimate of the relationships

between claims as if there had been no measurement error. The disattenuated correlation coefficients (R_{xy}) can be computed by using this formula:[7] $r_{xy}/(r_{xx} * r_{yy})^{0.5}$, where r_{xy} is the observed correlation and r_{xx} and r_{yy} are the reliabilities for Claim X and Claim Y. Given that none of the claims had perfect reliabilities, the disattenuated–claim correlations are higher than their observed–score counterparts.

[7] Spearman, 1904

Disattenuated correlations equal to or very near 1.00 suggest that the same construct is being measured. High values (e.g., low 0.90s) suggest that different claims are measuring either very similar constructs or different aspects of the same construct. Values markedly less than 1.00 suggest that the claims reflect different constructs.

None of the claims had perfect reliabilities (see the reliability chapter), the disattenuated==claim correlations are higher than their observed–score counterparts.

Some caution is needed when interpreting the disattenuated correlations because the reliabilities used to calculate these values are subject to both upward and downward biases. Consequently, some of the tabled values may be higher or lower than they should be, depending on which bias prevails for any given pair of claim scores. When the reliabilities are lower than they should be, the disattenuated correlations will be inflated (and in some instances can appear even larger than the theoretical correlation maximum value of 1.00).

These are discussed in some detail in the reliability chapter.

22.4.4 Claim-Score Correlations

Correlations between claim scores within each subject area are presented below. The KAP Mathematics tests have four claims. The KAP ELA tests have two claims; however, the first claim had enough items to report out individual scores for two sets of targets.

For each grade, Pearson's correlation coefficients between these claims are reported below. The correlations between the claims within the content areas are positive and generally moderate in value.

Table 22.15: Claim Score Correlations: Grade 3

	Math 1	Math 2	Math 3	Math 4	ELA 1 Info.	ELA 1 Lit.	ELA 2
Math 1	1.00	0.66	0.67	0.63	0.69	0.65	0.69
Math 2	0.66	1.00	0.52	0.49	0.52	0.49	0.51
Math 3	0.67	0.52	1.00	0.51	0.55	0.52	0.54
Math 4	0.63	0.49	0.51	1.00	0.53	0.50	0.52
ELA 1 Info.	0.69	0.52	0.55	0.53	1.00	0.76	0.77
ELA 1 Lit.	0.65	0.49	0.52	0.50	0.76	1.00	0.75
ELA 2	0.69	0.51	0.54	0.52	0.77	0.75	1.00

Table 22.16: Claim Score Correlations: Grade 4

	Math 1	Math 2	Math 3	Math 4	ELA 1 Info.	ELA 1 Lit.	ELA 2
Math 1	1.00	0.64	0.69	0.65	0.68	0.65	0.70
Math 2	0.64	1.00	0.51	0.50	0.49	0.47	0.50
Math 3	0.69	0.51	1.00	0.53	0.54	0.51	0.55
Math 4	0.65	0.50	0.53	1.00	0.50	0.47	0.50
ELA 1 Info.	0.68	0.49	0.54	0.50	1.00	0.75	0.75
ELA 1 Lit.	0.65	0.47	0.51	0.47	0.75	1.00	0.74
ELA 2	0.70	0.50	0.55	0.50	0.75	0.74	1.00

Table 22.17: Claim Score Correlations: Grade 5

	Math 1	Math 2	Math 3	Math 4	ELA 1 Info.	ELA 1 Lit.	ELA 2
Math 1	1.00	0.64	0.68	0.63	0.67	0.61	0.68
Math 2	0.64	1.00	0.52	0.48	0.48	0.43	0.48
Math 3	0.68	0.52	1.00	0.52	0.51	0.46	0.52
Math 4	0.63	0.48	0.52	1.00	0.48	0.43	0.49
ELA 1 Info.	0.67	0.48	0.51	0.48	1.00	0.72	0.76
ELA 1 Lit.	0.61	0.43	0.46	0.43	0.72	1.00	0.71
ELA 2	0.68	0.48	0.52	0.49	0.76	0.71	1.00

Table 22.18: Claim Score Correlations: Grade 6

	Math 1	Math 2	Math 3	Math 4	ELA 1 Info.	ELA 1 Lit.	ELA 2
Math 1	1.00	0.56	0.56	0.50	0.66	0.65	0.66
Math 2	0.56	1.00	0.39	0.34	0.42	0.41	0.42
Math 3	0.56	0.39	1.00	0.34	0.40	0.39	0.40
Math 4	0.50	0.34	0.34	1.00	0.38	0.38	0.38
ELA 1 Info.	0.66	0.42	0.40	0.38	1.00	0.75	0.74
ELA 1 Lit.	0.65	0.41	0.39	0.38	0.75	1.00	0.73
ELA 2	0.66	0.42	0.40	0.38	0.74	0.73	1.00

Table 22.19: Claim Score Correlations: Grade 7

	Math 1	Math 2	Math 3	Math 4	ELA 1 Info.	ELA 1 Lit.	ELA 2
Math 1	1.00	0.65	0.58	0.65	0.66	0.63	0.66
Math 2	0.65	1.00	0.45	0.50	0.46	0.43	0.46
Math 3	0.58	0.45	1.00	0.46	0.48	0.45	0.47
Math 4	0.65	0.50	0.46	1.00	0.52	0.50	0.52
ELA 1 Info.	0.66	0.46	0.48	0.52	1.00	0.74	0.74
ELA 1 Lit.	0.63	0.43	0.45	0.50	0.74	1.00	0.72
ELA 2	0.66	0.46	0.47	0.52	0.74	0.72	1.00

Table 22.20: Claim Score Correlations: Grade 8

	Math 1	Math 2	Math 3	Math 4	ELA 1 Info.	ELA 1 Lit.	ELA 2
Math 1	1.00	0.43	0.51	0.58	0.61	0.60	0.63
Math 2	0.43	1.00	0.28	0.28	0.26	0.24	0.26
Math 3	0.51	0.28	1.00	0.36	0.38	0.38	0.40
Math 4	0.58	0.28	0.36	1.00	0.48	0.48	0.48
ELA 1 Info.	0.61	0.26	0.38	0.48	1.00	0.73	0.72
ELA 1 Lit.	0.60	0.24	0.38	0.48	0.73	1.00	0.74
ELA 2	0.63	0.26	0.40	0.48	0.72	0.74	1.00

Table 22.21: Claim Score Correlations: Grade 10

	Math 1	Math 2	Math 3	Math 4	ELA 1 Info.	ELA 1 Lit.	ELA 2
Math 1	1.00	0.55	0.53	0.40	0.63	0.60	0.62
Math 2	0.55	1.00	0.35	0.29	0.42	0.41	0.41
Math 3	0.53	0.35	1.00	0.27	0.40	0.38	0.39
Math 4	0.40	0.29	0.27	1.00	0.29	0.27	0.27
ELA 1 Info.	0.63	0.42	0.40	0.29	1.00	0.74	0.73
ELA 1 Lit.	0.60	0.41	0.38	0.27	0.74	1.00	0.72
ELA 2	0.62	0.41	0.39	0.27	0.73	0.72	1.00

22.4.5 Exploratory Factor Analysis

Next year, additional studies will be done to explore the internal structure of the KAP assessments. An exploratory factor analysis (EFA) of the claim scores across all KAP subject areas can be conducted. Such a study will be attempted once the Science and HGSS tests become operational. Observed score correlation matrices can be used in these EFAs. AAI proposes using principle axis factor extraction with an oblique rotation (Promax) of the initial factor solution to improve interpretability. Oblique rotations allow for correlated factors, which seems more appropriate for KAP tests because of a priori expectations that academic achievement across subject areas should be correlated.

The internal structure evidence will be more complete after results from additional studies are available. IRT dimensionality studies currently support use of total scores to report student performance in math and ELA. Claim scores have less compelling evidence. Since the claims in each subject area were designed to measure its distinct components, it is reasonable to expect that the intersubject claim correlations should be positive, strong, and ideally, not extremely high. However, the evidence available so far imply that some claims are measuring essentially the same constructs.

There is less support for providing claim–score results beyond the total score. While content rationale may underlie the creation of claim scores, empirical correlations illustrate that caution is required when using claim scores to identify individual student strengths and weaknesses. Certainly, instructional programs should not be based on claim–score information alone but in conjunction with other sources of available evidence (e.g., teacher observations, other exam performances).

22.5 Evidence Based on Relationships with Other Variables

As described in the Standards:

> *Evidence based on relationships with other variables provides evidence about the degree to which these relationships are consistent with the construct underlying the proposed test score interpretations.*[8]

[8] AERA, APA, & NCME, 2014, p. 16

This category of evidence refers to external structure evidence and is classified on three types of evidence: convergent, discriminant, and criterion–related. Convergent evidence is provided by relationships between students' performance on different assessments intended to measure a similar construct. Discriminant evidence is provided by relationships between students' performance on different tests intended to measure different constructs. Criterion-related evidence, either predictive or concurrent, is provided by relationships between students' test scores and their performance on a criterion measure.[9]

[9] Cronbach, 1971; Messick, 1989

External validity evidence for the KAP tests has yet to be examined. Such studies could examine the correlations between KAP scores and a variety of other academic measures (e.g., ACT/SAT or other commonly administered assessments) to provide convergent and discriminant evidence. Criterion–related evidence could come from the relationships between the KAP and criterion variables such as grade point average (GPA), course grades, university proficiency exam scores, and students' GPA in their first college courses.

In addition, the relationship between the KAP and some irrelevant characteristics to determine whether the KAP exhibited any differential impact based on gender, ethnicity, English proficiency, or socioeconomic status. The differential item functioning (DIF) results presented in the fairness chapter address some of these concerns. Although some items had statistically significant differences, no items had effect-size differences that were practically significant.

The results from such studies would be expected to provide strong external evidence in support of the KAP as a valid measure of student achievement. Of course, empirical results are needed to confirm that hypothesis. Some limited discriminant validity evidence is available from comparing correlations across the KAP ELA and math tests. Each KAP assessment measures a different construct, so the correlations between ELA and math were not expected to be extremely high. The values in this table are consistent with this expectation as the correlations between the KAP ELA and math tests range from 0.69 to 0.76. Note that the correlations were fairly stable across the different grade levels and tended to decrease as grade level increased.

Grade	Correlation
3	0.76
4	0.75
5	0.74
6	0.73
7	0.73
8	0.70
10	0.69

Table 22.22: Correlation Between ELA and Math Scaled Scores

22.6 Evidence Based on the Consequences of Testing

Based on the Standards (2014), evidence of the consequences of implementing an assessment program is an additional source of validity information. Because the evaluation of consequential validity is so broadly defined, it can be difficult to measure specific aspects of consequential validity. Test data provide limited insight into this type of validation evidence.

The extent to which various groups of users (e.g., students, teachers, and parents) appropriately interpret these scores and reports appropriately can affect the validity of subsequent uses of these results. Separate chapters in this technical manual are dedicated to discussing the KAP item and test scores as well as score reports. Those chapters also provide information to help readers avoid unintended uses and interpretations of the KAP results. However, more evidence should be gathered regarding improved KAP test-score interpretation and decision-making.

22.6.1 Intended and Unintended Consequences

Both positive and negative (intended and unintended) consequences of score-based inferences must be investigated to fully evaluate the pool of validity evidence. The consequences of an assessment program alone do not serve as indicators of validity. It is the investigation and evaluation of the consequences that provide a richer context for establishing the validity of an assessment program. Because this was the KAP program's first year, some consequences may take longer to become evident. One intended consequence is eventual longitudinal improvement in KAP test scores. A common unintended consequence, seen in other testing programs, is an increase in cheating.

22.7 Evidence Related to the Use of IRT

Because IRT is the basis of all calibration, scaling, and linking analyses associated with the KAP, the validity of inferences from these results depends on the degree to which the assumptions of the IRT model are met, as well as the goodness of fit between the model and test data. As discussed at length in the IRT chapter, the underlying assumptions of IRT models were essentially met for all KAP data, indicating the appropriateness of using IRT models to analyze KAP data. However, additional studies will continue to expand the evidence base on the use of IRT.

In addition, the IRT model was used to link different operational KAP tests across years. Linking accuracy also affects the accuracy of student scores and the validity of score uses. As described in the linking chapter, only within-year linking of scores from different forms was required this year. Calibration was undertaken on randomly equivalent samples of students. However, a set of common items also appeared on all forms used for any particular grade-level, subject-area test. This particular linking design was very robust. Next year, test–score linking will be more involved than this year's linking because across–year linking will also be required.

22.8 Evidence Related to Standard Setting

See the standard setting chapter for more information.

Following Round 3 at the Bookmark standard-setting event, panelists completed a *Final Evaluation Form*, where most questions required panelists to respond on a Likert-like scale ranging from *Strongly Disagree* (1) to *Strongly Agree* (6).

The primary purpose of several of the items was to gather validity evidence pertaining to panelists' confidence and comfort with the performance levels (Level 2, Level 3, and Level 4), and with the final cut score for each of the three levels. There were three questions for each cut score. These were:

- The impact result (i.e., percentage of students) for this achievement level is reasonable,
- The cut score for this achievement level is appropriate based on the PLDs and the just-barely student activities, and
- The cut score for this achievement is defensible due to panelistsâ adherence to procedure.

The mean ratings for the three questions across the three cut scores were generally high, with more than 70% of the ratings being 5 or 6 on the Likert scale for every question. The ratings for ELA Grade 3 were generally the lowest.

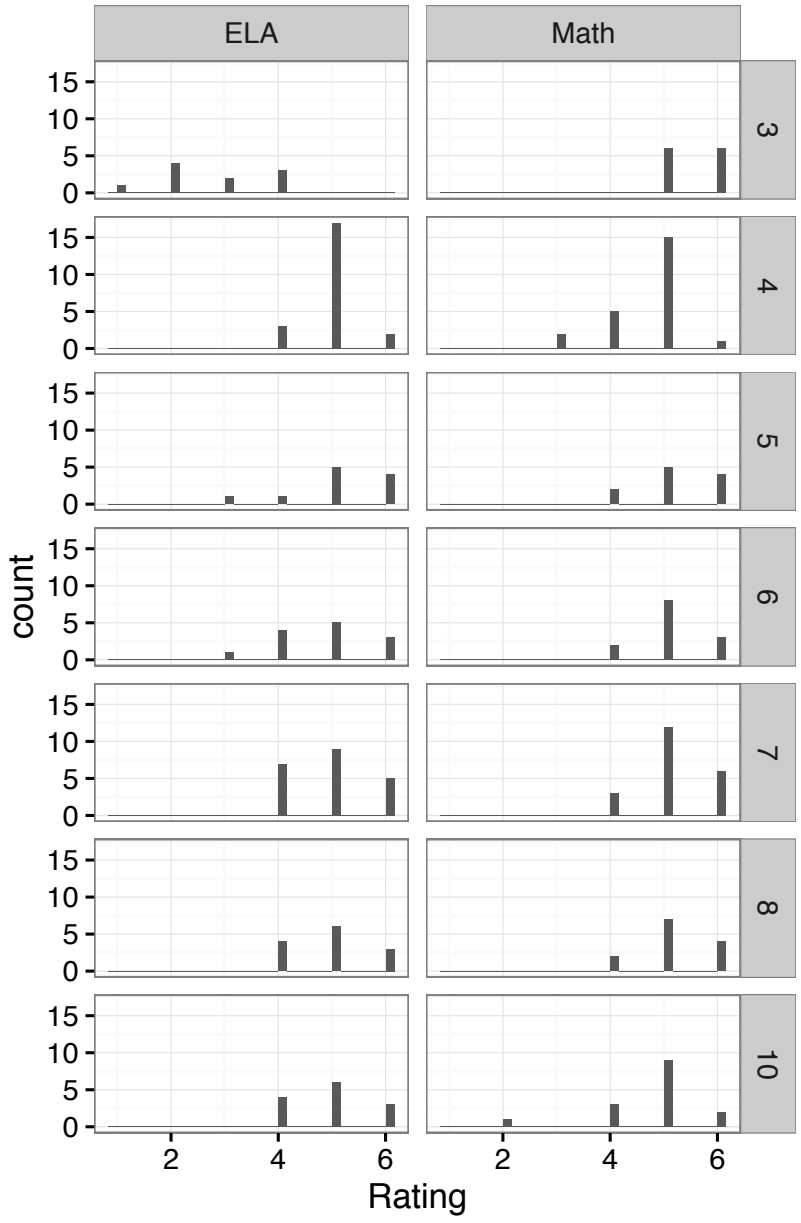


Figure 22.1: Panelists' Ratings for the Reasonableness of the Level 2 Cut Score Based on the Impact Results

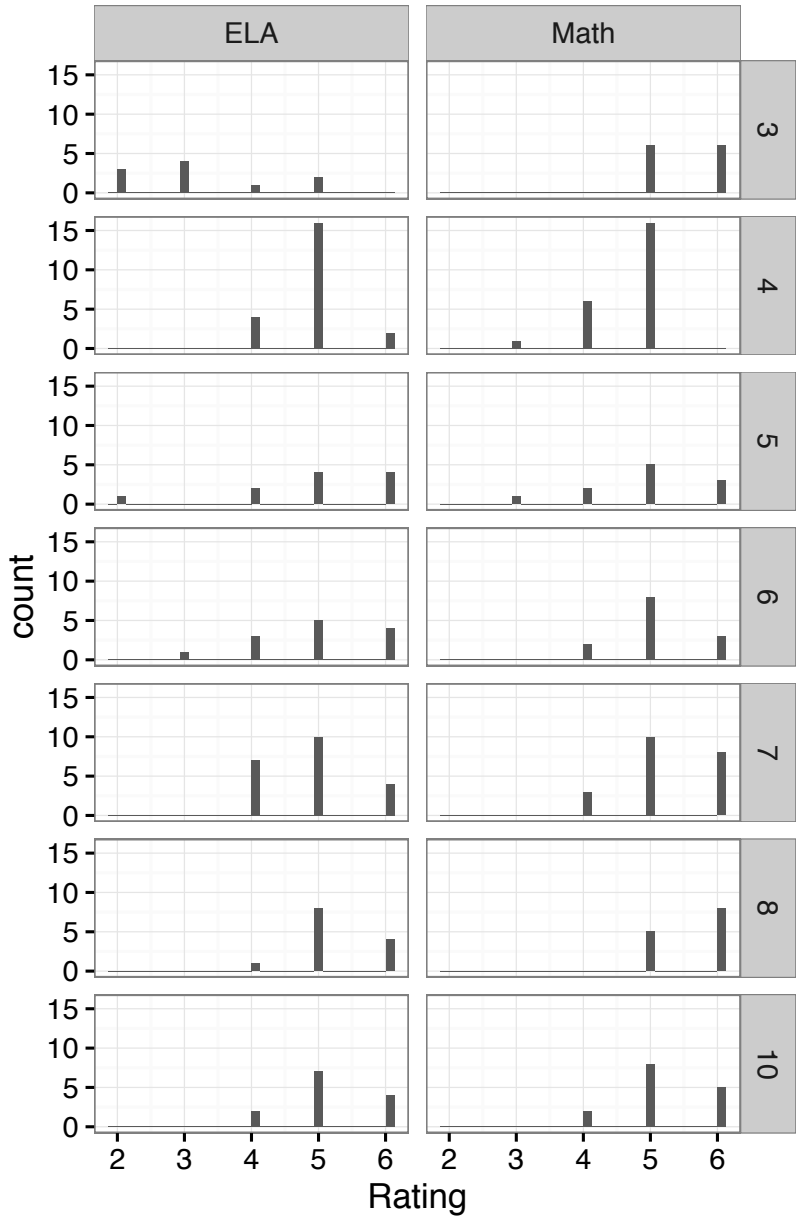


Figure 22.2: Panelists' Ratings for the Appropriateness of the Level 2 Cut Score Based on the PLDs and Just-Barely Student Activities

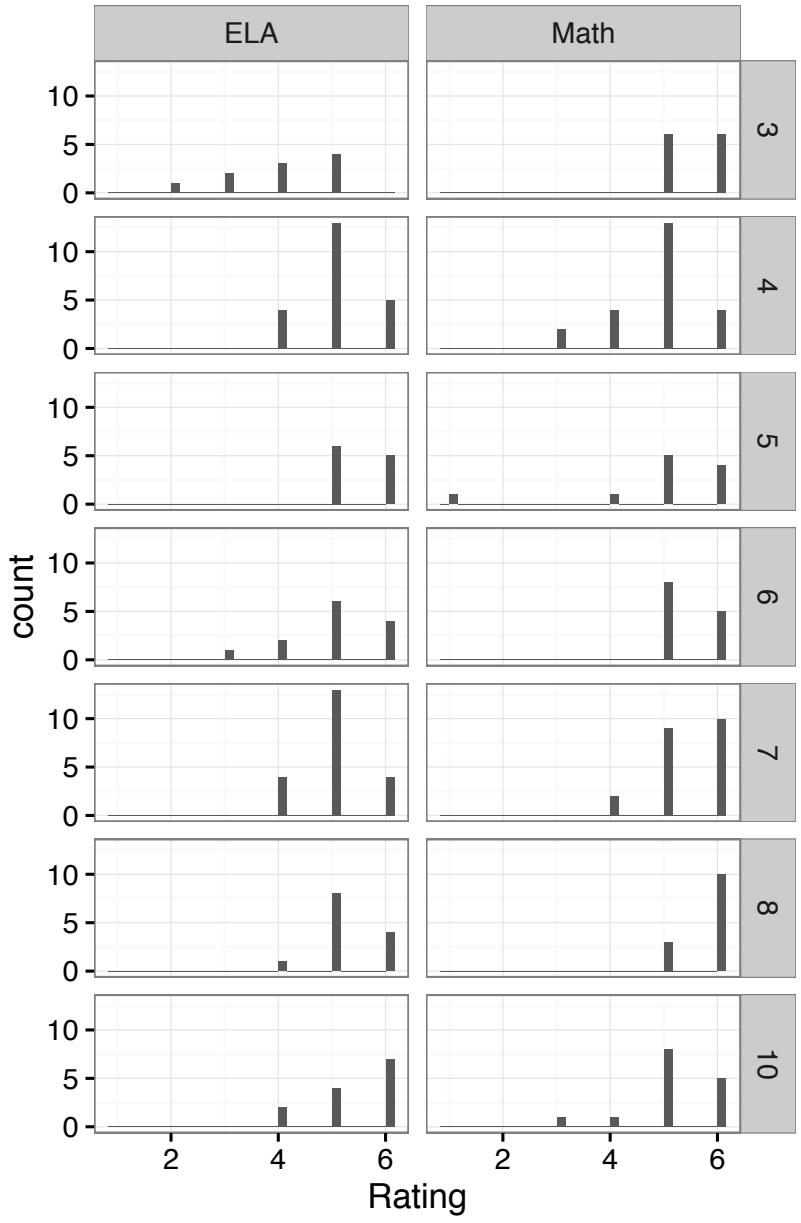


Figure 22.3: Panelists' Ratings for the Defensibleness of the Level 2 Cut Score Based on the Panelist Adherence to Procedures

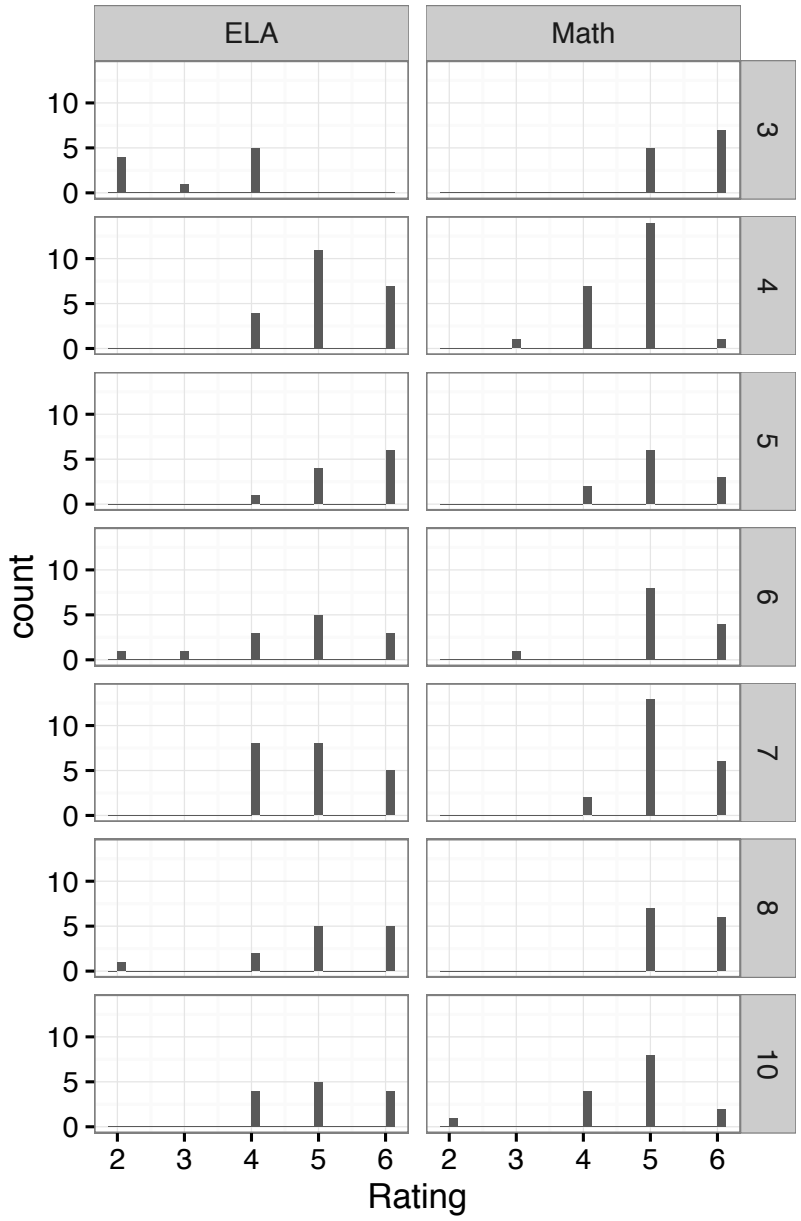


Figure 22.4: Panelists' Ratings for the Reasonableness of the Level 3 Cut Score Based on the Impact Results

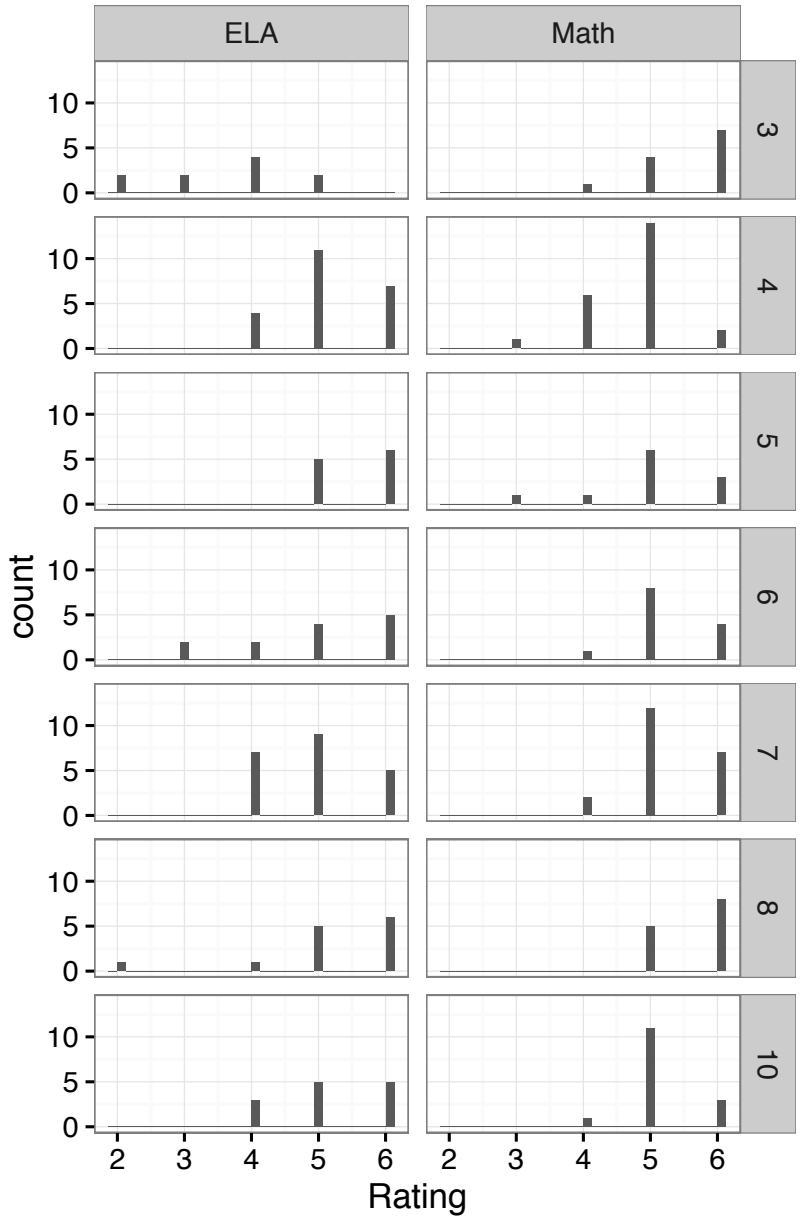


Figure 22.5: Panelists' Ratings for the Appropriateness of the Level 3 Cut Score Based on the PLDs and Just-Barely Student Activities

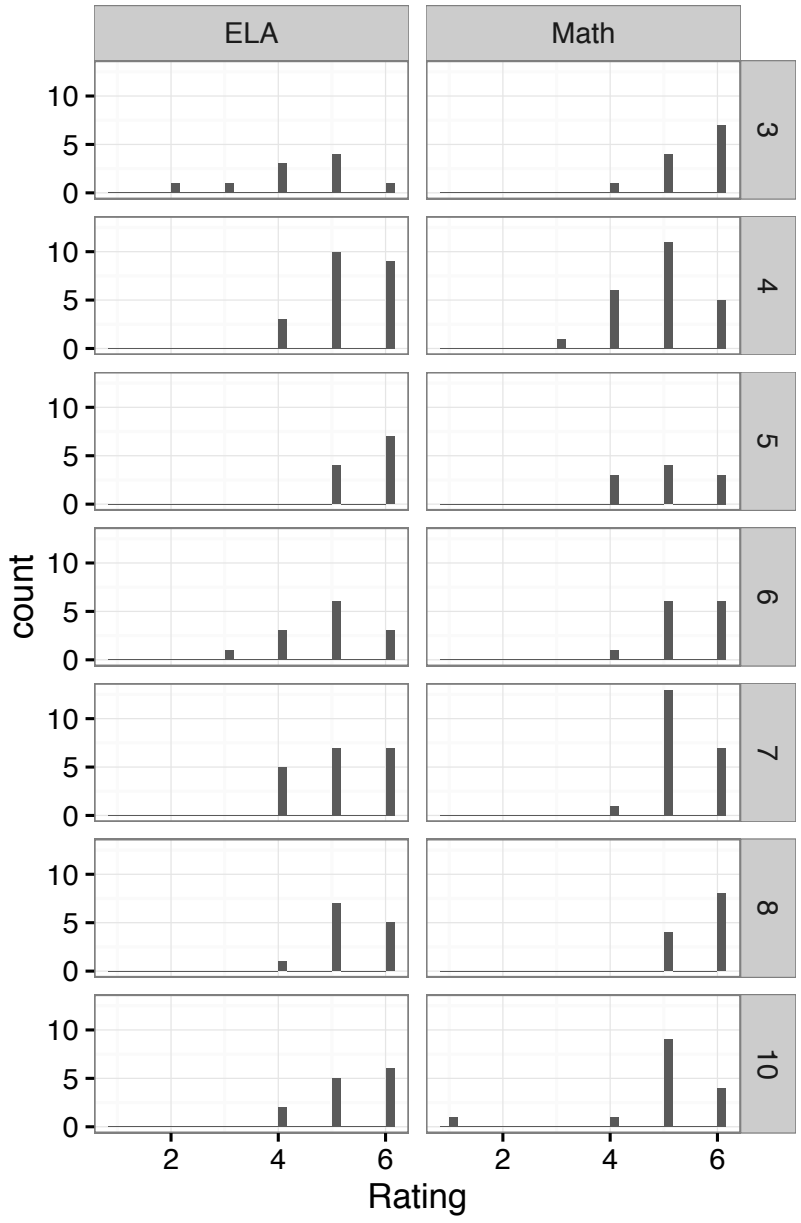


Figure 22.6: Panelists' Ratings for the Defensibleness of the Level 3 Cut Score Based on the Panelist Adherence to Procedures

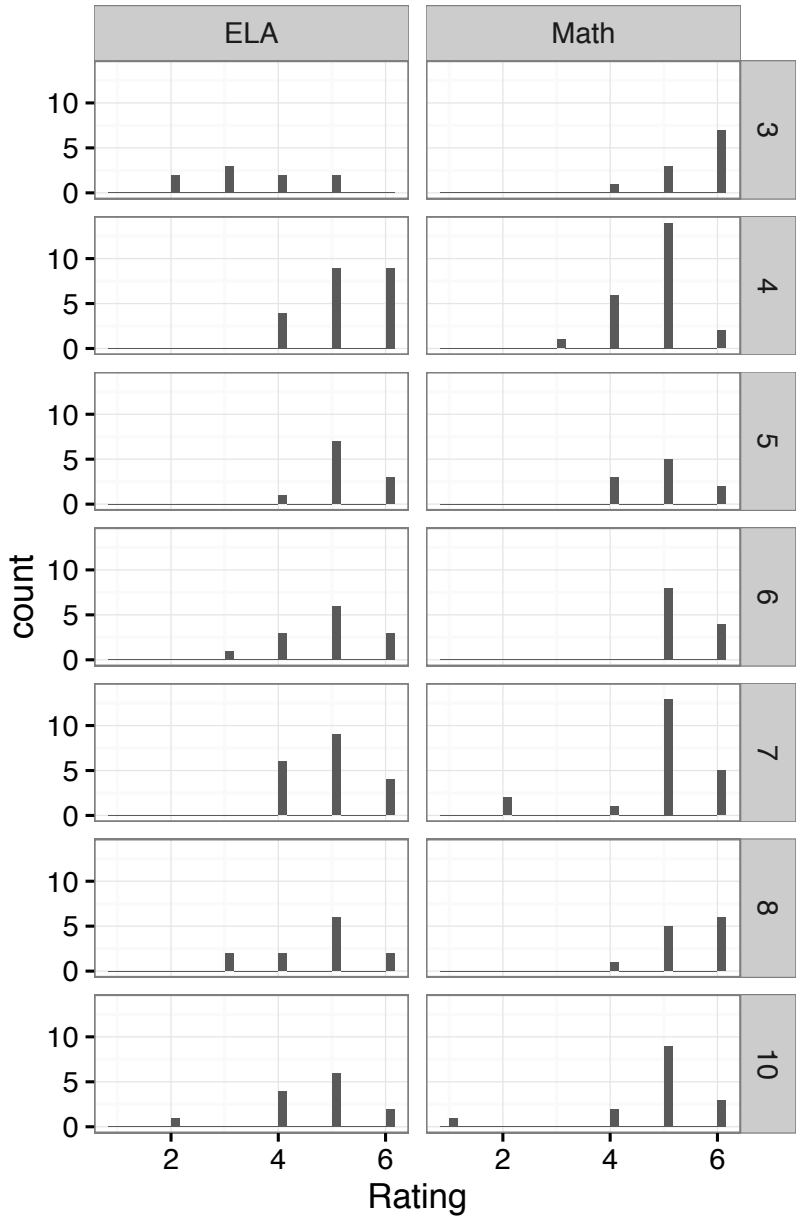


Figure 22.7: Panelists' Ratings for the Reasonableness of the Level 4 Cut Score Based on the Impact Results

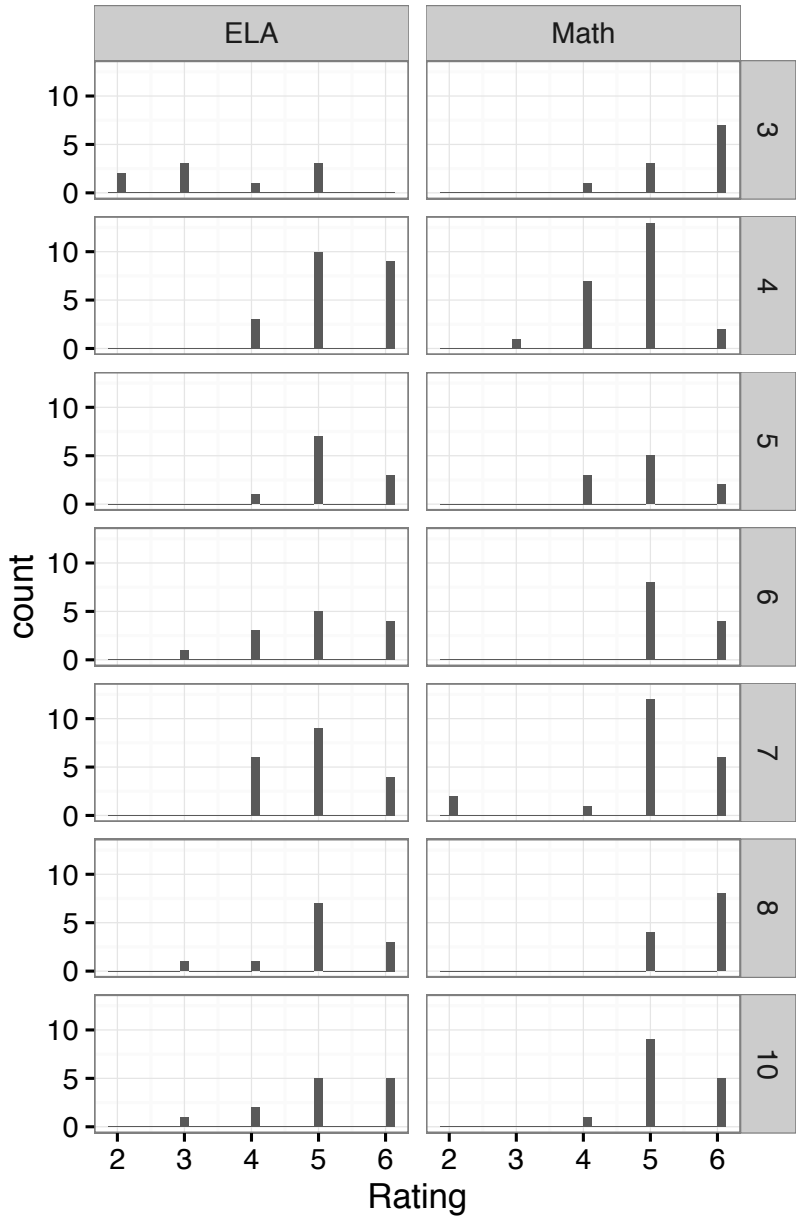


Figure 22.8: Panelists' Ratings for the Appropriateness of the Level 4 Cut Score Based on the PLDs and Just-Barely Student Activities

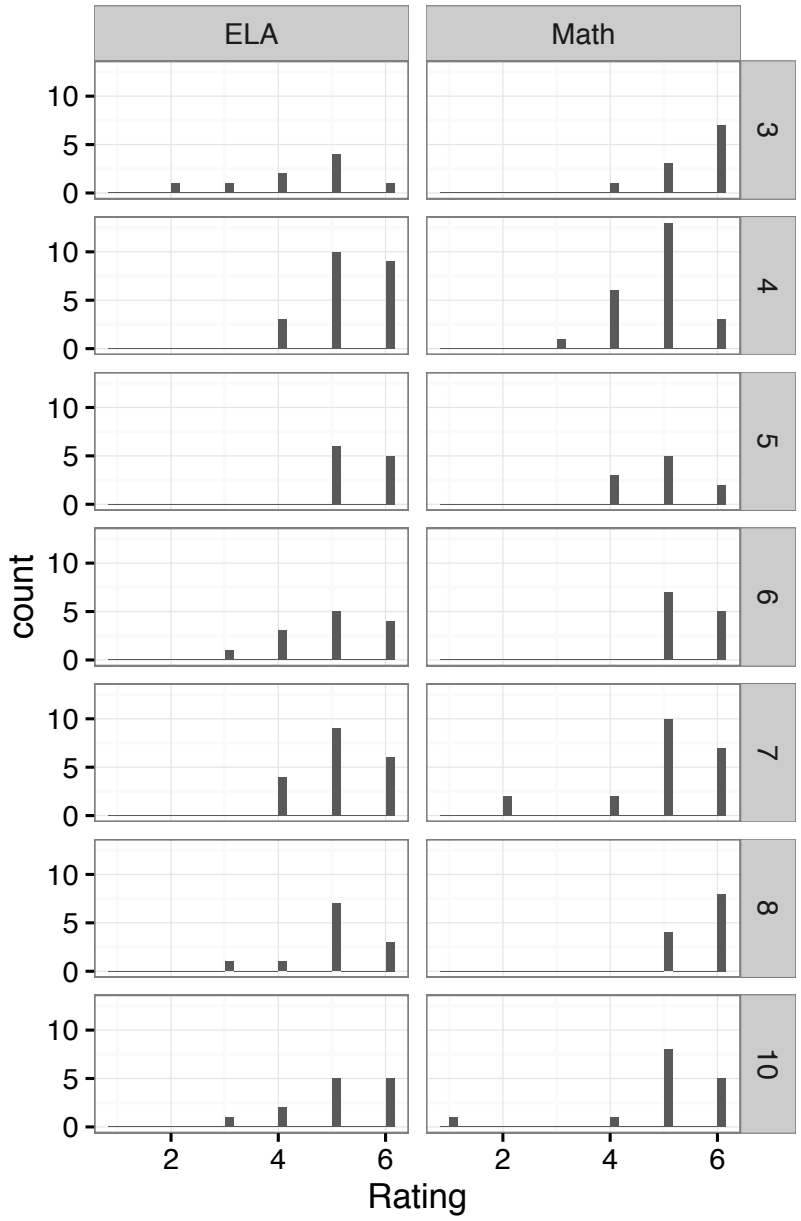


Figure 22.9: Panelists' Ratings for the Defensibleness of the Level 4 Cut Score Based on the Panelist Adherence to Procedures

22.9 Validity Evidence Summary

Validity evidence related to test content was reviewed earlier in this chapter. On the whole, the early chapters of this technical report show a link between each KAP item and the KCCRS. The chapters on item and test development presented details about how

KAP operational assessments were created to reflect the KCCRS, as well as comprehensive information about educator reviews, including content, bias, and sensitivity reviews.

Test– and claim–score intercorrelations were presented earlier in this chapter. In general, within-subject-area claims (e.g., mathematics) correlate more highly with themselves than they do with other subject-area claims (e.g., ELA) providing favorable evidence for the internal and external relationships between test components.

In math, the correlation between Claim 1 scores and the combined scores from Claims 2, 3, and 4 were considered for the within-subject relationship. Those correlations were provided in the section on the added value of subscores. The item counts for math Claims 2, 3, and 4 were small enough to attenuate those correlations to the point that a fair comparison with within-subject ELA correlations was not possible.

Validity of score inferences is bolstered when test scores are consistent. Here, the reliabilities of the total test scores (see the reliability chapter) are very good, with many being in the low 0.90s.

Additionally, as reported in the fairness chapter, differential item functioning (DIF) with respect to gender and ethnicity helps address construct-irrelevant variance, which presents a serious threat to the validity of inferences made from achievement test scores. As noted in that chapter, items were screened and reviewed for DIF.

Because most students took the KAP online using the KITE testing engine, testing mode artifacts pose a minimal threat to the validity of KAP test scores. Other administration modes exist but are very rarely used. Although alternate forms were used, a strong linking design reduced the likelihood of threats from their usage. The use of multiple forms likely reduced the risk of students copying each other's answers during testing. With any achievement test, a significant threat can arise from alignment issues between test items and the curriculum used with students. The use of industry standard test–development procedures almost certainly mitigated this threat. Results from independent reviews will be available soon.

As noted in the linking chapter subtle content differences across forms likely prevent interpreting the linking results as equating.

22.9.1 Overview of Future Validity Studies

The need for future research has already been stated. TAC and peer-reviewer feedback likely will guide some of this research. KSDE and AAI welcome suggestions for research leading to improvement of the KAP and its validity. Planned research projects include:

- Expanded IRT model evaluation. Each technical manual will explore one or two special research issues in depth. This year, item invariance was a focus. Next year's manual will focus on invariance over students.
- Exploratory factor analysis (EFA) of KAP claim scores. EFA will be studied once HGSS and Science assessments are operational.
- Alignment work by edCount LLC. This work will be available next year.
- Suitable external criterion measures (e.g., college admission test scores). After these measures are available, AAI will work with

KSDE and its TAC to plan validity studies.

Part VII

Appendices

A
Math Content Emphasis

Figure A.1: Content Emphases for Math Page 1

Mathematics Content Emphases

The pattern of emphasis for the Targets that compose the Claims is adapted from the work of national assessment initiatives. Individual standards, while important, are impossible to accurately measure with limited testing time. By assessing at the Target level, it is possible to highlight student comprehension of the connected material contained in the Standards. To capture the focus, coherence, and rigor of the Standards, it is necessary to vary the emphasis on particular Targets. All of the content is eligible for assessment, and the balance of tested content is derived from the expectations of the Standards.

The Claims are the broadest categories of knowledge, skills, and abilities that can have inferences drawn about them. Claims are built from Targets; Targets are drawn from the Standards.

The Goal Depth of Knowledge (an index of cognitive complexity) is provided as a general reference for the projected maximum DOK of items. Typically, items are at DOK 1 or 2, with some DOK 3 items as supported by the context. DOK 4 is generally reserved for performance tasks, such as geometric proofs or figure constructions.

The Relative Emphasis for each Target in Claim 1 is based on the work of the national assessment initiatives and the relative frequency with which items aligned to that Target would appear on an item-adaptive test. The Relative Emphasis should **NOT** be interpreted as a basis for making curricular decisions. Targets with a Low Relative Emphasis may include concomitant skills of other Medium or High Targets in the same grade. These Targets may also be important foundational skills in a progression, and key to success in later grades.

Figure A.2: Content Emphases for Math Page 2

Content Emphases for Grade 3

Claim (% of Test)	Target(s)	Goal DOK	Relative Emphasis/Comments
1. Concepts & Procedures (65-75%)	A	2	High
	B	1	High
	C	1	High
	D	2	High
	E	1	Low
	F	2	High
	G	2	High
	H	3	Medium
	I	2	High
	J	2	Low
	K	2	Medium
2. Problem Solving (8-12%)	A–D	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.
3. Communicating Reasoning (8-12%)	A–F	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.
4. Modeling and Data Analysis (8-12%)	A–G	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.

Figure A.3: Content Emphases for Math Page 3

Grade 3, Claim 1 Targets

Target A	Represent and solve problems involving multiplication and division.
Target B	Understand properties of multiplication and the relationship between multiplication and division.
Target C	Multiply and divide up to 100.
Target D	Solve problems involving the four operations, and identify and explain patterns in arithmetic.
Target E	Use place value understanding and properties of operations to perform multi-digit arithmetic.
Target F	Develop understanding of fractions as numbers.
Target G	Solve problems involving measurement and estimation of intervals of time, liquid volumes, and masses of objects.
Target H	Represent and interpret data.
Target I	Geometric measurement: understand concepts of area and relate area to multiplication and to addition.
Target J	Geometric measurement: recognize perimeter as an attribute of plane figures and distinguish between linear and area measures.
Target K	Reason with shapes and their attributes.

Figure A.4: Content Emphases for Math Page 4

Content Emphases for Grade 4

Claim (% of Test)	Target(s)	Goal DOK	Relative Emphasis/Comments
1. Concepts & Procedures (65-75%)	A	2	High
	B	1	Medium
	C	3	Low
	D	2	High
	E	2	High
	F	2	High
	G	2	High
	H	2	High
	I	2	Medium
	J	2	Medium
	K	2	Low
	L	2	Low
2. Problem Solving (8-12%)	A–D	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.
3. Communicating Reasoning (8-12%)	A–F	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.
4. Modeling and Data Analysis (8-12%)	A–G	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.

Figure A.5: Content Emphases for Math Page 5

Grade 4, Claim 1 Targets

Target A	Use the four operations with whole numbers to solve problems.
Target B	Gain familiarity with factors and multiples.
Target C	Generate and analyze patterns.
Target D	Generalize place value understanding for multi-digit whole numbers.
Target E	Use place value understanding and properties of operations to perform multi-digit arithmetic.
Target F	Extend understanding of fraction equivalence and ordering.
Target G	Build fractions from unit fractions by applying and extending previous understandings of operations on whole numbers.
Target H	Understand decimal notation for fractions, and compare decimal fractions.
Target I	Solve problems involving measurement and conversion of measurements from a larger unit to a smaller unit, and involving time.
Target J	Represent and interpret data.
Target K	Geometric measurement: understand concepts of angle and measure angles.
Target L	Draw and identify lines and angles, and classify shapes by properties of their lines and angles.

Figure A.6: Content Emphases for Math Page 6

Content Emphases for Grade 5

Claim (% of Test)	Target(s)	Goal DOK	Relative Emphasis/Comments
1. Concepts & Procedures (65-75%)	A	1	Low
	B	2	Low
	C	2	High
	D	2	High
	E	2	High
	F	2	High
	G	1	Medium
	H	2	Medium
	I	2	High
	J	1	Low
	K	2	Low
2. Problem Solving (8-12%)	A–D	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.
3. Communicating Reasoning (8-12%)	A–F	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.
4. Modeling and Data Analysis (8-12%)	A–G	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.

Figure A.7: Content Emphases for Math Page 7

Grade 5, Claim 1 Targets

Target A	Write and interpret numerical expressions.
Target B	Analyze patterns and relationships.
Target C	Understand the place value system.
Target D	Perform operations with multi-digit whole numbers and with decimals to hundredths.
Target E	Use equivalent fractions as a strategy to add and subtract fractions.
Target F	Apply and extend previous understandings of multiplication and division to multiply and divide fractions.
Target G	Convert like measurement units within a given measurement system and solve problems involving time.
Target H	Represent and interpret data.
Target I	Geometric measurement: understand concepts of volume and relate volume to multiplication and to addition.
Target J	Graph points on the coordinate plane to solve real-world and mathematical problems.
Target K	Classify two-dimensional (plane) figures into categories based on their properties.

Figure A.8: Content Emphases for Math Page 8

Content Emphases for Grade 6

Claim (% of Test)	Target(s)	Goal DOK	Relative Emphasis/Comments
1. Concepts & Procedures (65-75%)	A	2	High
	B	2	High
	C	1	Low
	D	2	High
	E	2	High
	F	2	High
	G	2	High
	H	2	Medium
	I	2	Low
	J	2	Low
2. Problem Solving (8-12%)	A–D	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.
3. Communicating Reasoning (8-12%)	A–F	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.
4. Modeling and Data Analysis (8-12%)	A–G	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.

Figure A.9: Content Emphases for Math Page 9

Grade 6, Claim 1 Targets

Target A	Understand ratio concepts and use ratio reasoning to solve problems.
Target B	Apply and extend previous understandings of multiplication and division to divide fractions by fractions.
Target C	Compute fluently with multi-digit numbers and find common factors and multiples.
Target D	Apply and extend previous understandings of numbers to the system of rational numbers.
Target E	Apply and extend previous understandings of arithmetic to algebraic expressions.
Target F	Reason about and solve one-variable equations and inequalities.
Target G	Represent and analyze quantitative relationships between dependent and independent variables.
Target H	Solve real-world and mathematical problems involving area, surface area, and volume.
Target I	Develop an understanding of statistics variability.
Target J	Summarize and describe distributions.

Figure A.10: Content Emphases for Math Page 10

Content Emphases for Grade 7

Claim (% of Test)	Target(s)	Goal DOK	Relative Emphasis/Comments
1. Concepts & Procedures (65-75%)	A	2	High
	B	2	High
	C	1	High
	D	2	High
	E	3	Low
	F	2	Low
	G	2	Medium
	H	2	Low
	I	2	Medium
2. Problem Solving (8-12%)	A–D	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.
3. Communicating Reasoning (8-12%)	A–F	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.
4. Modeling and Data Analysis (8-12%)	A–G	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.

Grade 7, Claim 1 Targets

Target A	Analyze proportional relationships and use them to solve real-world and mathematical problems.
Target B	Apply and extend previous understandings of operations with fractions to add, subtract, multiply, and divide rational numbers.
Target C	Use properties of operations to generate equivalent expressions.
Target D	Solve real-life and mathematical problems using numerical and algebraic expressions and equations.
Target E	Draw, construct, and describe geometrical figures and describe the relationships between them.
Target F	Solve real-life and mathematical problems involving angle measure, area, surface area, and volume.
Target G	Use random sampling to draw inferences about a population.
Target H	Draw informal comparative inferences about two populations.
Target I	Investigate chance processes and develop, use, and evaluate probability models.

Figure A.11: Content Emphases for Math Page 11

Content Emphases for Grade 8

Claim (% of Test)	Target(s)	Goal DOK	Relative Emphasis/Comments
1. Concepts & Procedures (65-75%)	A	1	Medium
	B	1	High
	C	2	High
	D	2	High
	E	2	High
	F	2	Medium
	G	2	High
	H	2	High
	I	2	Low
	J	2	Medium
2. Problem Solving (8-12%)	A–D	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.
3. Communicating Reasoning (8-12%)	A–F	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.
4. Modeling and Data Analysis (8-12%)	A–G	3	Tasks limited to machine-scorable responses, so not all Targets may be addressed.

Grade 8, Claim 1 Targets

Target A	Know that there are numbers that are not rational, and approximate them by rational numbers.
Target B	Work with radicals and integer exponents.
Target C	Understand the connections between proportional relationships, lines, and linear equations.
Target D	Analyze and solve linear equations and pairs of simultaneous linear equations.
Target E	Define, evaluate, and compare functions.
Target F	Use functions to model relationships between quantities.
Target G	Understand congruence and similarity using physical models, transparencies, or geometry software.
Target H	Understand and apply the Pythagorean Theorem.
Target I	Solve real-world and mathematical problems involving volume of cylinders, cones, and spheres.
Target J	Investigate patterns of association in bivariate data.

Figure A.12: Content Emphases for Math Page 12

Content Emphases for Grade 10

Claim (% of Test)	Target(s)	Goal DOK	Relative Emphasis/Comments
1. Concepts & Procedures (65-75%)	A	2	Low
	C	2	Medium
	D	1	Medium
	E	2	Medium
	F	1	Medium
	G	2	High
	H	2	Medium-High
	I	2	Medium
	J	2	Very High
	K	1	Medium
	L	2	Medium-High
	M	3	Medium
	N	2	Low
	O	2	Low
	P	2	Medium
	Q	2	High
	R	2	Medium
2. Problem Solving (8-12%)	A–D	3	Tasks are limited to machine-scorable responses, so not all Targets will be addressed.
3. Communicating Reasoning (8-12%)	A–F	3	Tasks are limited to machine-scorable responses, so not all Targets will be addressed.
4. Modeling and Data Analysis (8-12%)	A–G	3	Tasks are limited to machine-scorable responses, so not all Targets will be addressed.

Figure A.13: Content Emphases for Math Page 13

Grade 10, Claim 1 Targets

Target A	Extend the properties of exponents to rational exponents.
Target C	Reason quantitatively and use units to solve problems.
Target D	Interpret the structure of expressions.
Target E	Write expressions in equivalent forms to solve problems.
Target F	Perform arithmetic operations on polynomials.
Target G	Create equations that describe numbers or relationships.
Target H	Understand solving equations as a process of reasoning and explain the reasoning.
Target I	Solve equations and inequalities in one variable.
Target J	Represent and solve equations and inequalities graphically.
Target K	Understand the concept of a function and use function notation.
Target L	Interpret functions that arise in applications in terms of the context.
Target M	Analyze functions using different representations.
Target N	Build a function that models a relationship between two quantities.
Target O	Define trigonometric ratios and solve problems involving right triangles.
Target P	Summarize, represent, and interpret data on a single count or measurement variable.
Target Q	Prove geometric theorems.
Target R	Explain volume formulas and use them to solve problems.

Figure A.14: Content Emphases for Math Page 14

Claims 2, 3, and 4 Targets – All Grades

Claim 2: Problem Solving

Target A	Apply mathematics to solve well-posed problems in pure mathematics and arising in everyday life, society, and the workplace.
Target B	Select and use appropriate tools strategically.
Target C	Interpret results in the context of a situation.
Target D	Identify important quantities in a practical situation and map their relationships (e.g., using diagrams, two-way tables, graphs, flowcharts, or formulas).

Claim 3: Communicating Reasoning

Target A	Test propositions or conjectures with specific examples.
Target B	Construct, autonomously, chains of reasoning that will justify or refute propositions or conjectures.
Target C	State logical assumptions being used.
Target D	Use the technique of breaking an argument into cases.
Target E	Distinguish correct logic or reasoning from that which is flawed and—if there is a flaw in the argument— explain what it is.
Target F	Base arguments on concrete referents such as objects, drawings, diagrams, and actions.

Claim 4: Modeling and Data Analysis

Target A	Apply mathematics to solve problems arising in everyday life, society, and the workplace.
Target B	Construct, autonomously, chains of reasoning to justify mathematical models used, interpretations made, and solutions proposed for a complex problem.
Target C	State logical assumptions being used.
Target D	Interpret results in the context of a situation.
Target E	Analyze the adequacy of and make improvements to an existing model or develop a mathematical model of a real phenomenon.

Figure A.15: Content Emphases for Math Page 15

Target F	Identify important quantities in a practical situation and map their relationships (e.g., using diagrams, two-way tables, graphs, flowcharts, or formulas).

B
ELA Content Emphasis

Figure B.1: Content Emphases for ELA Page 1

ELA Content Emphases

The pattern of emphasis for the Targets that compose the Claims is adapted from the work of national assessment initiatives. Individual standards, while important, are impossible to accurately measure with limited testing time. By assessing the Target level, it is possible to highlight student comprehension of the connected material contained in the Standards. To capture the focus, coherence, and rigor of the Standards, it is necessary to vary the emphasis on particular Targets. All of the content is eligible for assessment, and the balance of tested content is derived from the expectations of the Standards.

The Claims are the broadest categories of knowledge, skills, and abilities that can have inferences drawn about them. In ELA, each Claim contains one or more sections that each indicate a Focus on a particular skill or area of the larger content. Claims are built from Targets; Targets are drawn from the Standards. The distribution of texts and items is relatively equal among the foci for each claim.

The Goal Depth of Knowledge (an index of cognitive complexity) is provided as a general reference for the projected maximum DOK of items. Typically, items are at DOK 1 or 2, with some DOK 3 items as supported by the text. DOK 4 is generally reserved for performance tasks, such as composition.

The Relative Emphasis for each Target is based on the work of the national assessment initiatives and the relative frequency with which items aligned to that Target would appear on an item-adaptive test. The Relative Emphasis should **NOT** be interpreted as a basis for making curricular decisions. Targets with a Low Relative Emphasis may include concomitant skills of other Medium or High Targets in the same grade, or they may be important building-block skills and are key to success in later grades.

Figure B.2: Content Emphases for ELA Page 2

Content Emphases for Grades 3-5

Claim (% of Test)	Focus	Target	Goal DOK	Relative Emphasis
1. Reading (60-65%)	Literary Texts	1: Key Details	2	Medium
		2: Central Ideas	2	High
		3: Word Meanings	2	Medium
		4: Reasoning & Evidence	3	High
		5: Analysis Within Or Across Texts	3	Low
		6: Text Structures & Features	3	
		7: Language Use	3	
	Informational Texts	8: Key Details	2	Medium
		9: Central Ideas	2	High
		10: Word Meanings	2	Medium
		11: Reasoning & Evidence	3	High
		12: Analysis Within Or Across Texts	3	Low
		13: Text Structures & Features	3	
		14: Language Use	3	
2. Writing (25-30%)	Write / Revise	1/3/6: Write / Revise Brief Texts	2	High
	Language / Vocabulary	8: Language & Vocabulary Use	1	High
	Conventions	9: Edit	1	High
3. Listening (10-15%)	Listen	4: Listen / Interpret	3	High

Figure B.3: Content Emphases for ELA Page 3

Grade 3, Claim 1 Targets

Target 1	Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.
Target 2	Identify or summarize central ideas, key events, the sequence of events, or the author's message or purpose presented in a text.
Target 3	Determine intended meanings of words, including multiple meanings of academic/tier 2 words, based on context, word relationships, word structure (e.g., common roots, affixes), or use of resources (e.g., beginning dictionary), with primary focus on determining meaning based on context and the academic/tier 2 vocabulary common to complex texts in all disciplines.
Target 4	Make an inference or provide a conclusion and use supporting evidence to justify/explain inferences (e.g., character development/actions/traits; first- or third-person point of view; theme; author's message or purpose).
Target 5	Examine or compare relationships (literary elements: setting, conflict, dialogue, point of view, characterization) within or across texts.
Target 6	Relate knowledge of text structures, genre-specific features, or formats (visual/graphic/auditory effects) to obtain, interpret, explain, or connect information within text.
Target 7	Interpret use of language by distinguishing literal from non-literal meanings of words and phrases used in context.
Target 8	Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.
Target 9	Identify central ideas, key events, or procedures and details that support them.
Target 10	Determine intended meanings of words, including academic/tier 2 words, domain-specific/tier 3 words, and words with multiple meanings, based on context, word relationships (e.g., synonyms), word structure (e.g., common Greek or Latin roots, affixes), or use of resources (e.g., dictionary, glossary), with primary focus on the academic vocabulary common to complex texts in all disciplines.
Target 11	Make an inference or provide a conclusion and use supporting evidence to justify/explain inferences (e.g., author's line of reasoning, point of view/purpose, relevance of evidence or elaboration to support claims, concepts, ideas).
Target 12	Examine, integrate, or compare information or presentation of information within or across texts (e.g., cause and effect, integrate information).
Target 13	Relate knowledge of text structures or text features (e.g., graphics, bold text, headings) to obtain, interpret, or explain information.

Figure B.4: Content Emphases for ELA Page 4

Target 14	Interpret use of language by distinguishing literal from nonliteral meanings of words and phrases used in context

Grade 3, Claim 2 Targets

Target 1a	Demonstrate ability to use specific narrative techniques (e.g., dialogue, description), chronology, appropriate transitional strategies for coherence, or authors' craft appropriate to purpose (e.g., closure, detailing characters, plot, setting, an event).
Target 1b	Revise one or more paragraphs demonstrating specific narrative techniques (e.g., dialogue, description), chronology, appropriate transitional strategies for coherence, or authors' craft.
Target 3a	Demonstrate ability to organize ideas in informational/explanatory texts by stating a focus (main idea), including appropriate transitional strategies for coherence, or supporting details, or an appropriate conclusion.
Target 3b	Revise one or more informational/explanatory paragraphs demonstrating ability to organize ideas by stating a focus (main idea), including appropriate transitional strategies for coherence, or supporting details, or an appropriate conclusion.
Target 6a	Demonstrate ability to state opinions about topics or sources; set a context, organize ideas, develop supporting reasons, or provide an appropriate conclusion.
Target 6b	Revise one or more paragraphs demonstrating ability to state opinions about topics or sources; set a context, organize ideas, develop supporting reasons, or provide an appropriate conclusion.
Target 8	Accurately use language and vocabulary (including academic and domain-specific vocabulary) appropriate to the purpose and audience when revising or composing texts.
Target 9	Apply or edit grade-appropriate grammar usage, capitalization, punctuation, and spelling to clarify a message and edit narrative, explanatory/informational, and opinion texts.

Grade 3, Claim 3 Target

Target 4	Interpret and use information delivered orally.

Figure B.5: Content Emphases for ELA Page 5

Grade 4, Claim 1 Targets

Target 1	Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.
Target 2	Identify or summarize central ideas, key events, the sequence of events, or the author's message or purpose presented in a text.
Target 3	Determine intended meanings of words, including multiple meanings of academic/tier 2 words, based on context, word relationships (e.g., synonyms), word structure (e.g., common Greek or Latin roots, affixes), or use of resources (e.g., dictionary, thesaurus), with primary focus on determining meaning based on context and the academic/tier 2 vocabulary common to complex texts in all disciplines.
Target 4	Make an inference or provide a conclusion and use supporting evidence to justify/explain inferences (e.g., character development/actions/traits; first- or third-person point of view; theme; author's message or purpose).
Target 5	Examine or compare relationships (literary elements: setting, conflict, dialogue, point of view, characterization) within or across texts.
Target 6	Relate knowledge of text structures, genre-specific features, or formats (visual/graphic/auditory effects) to obtain, interpret, explain, or connect information within text.
Target 7	Interpret figurative language, literary devices, or connotative meanings of words and phrases used in context and the impact of those word choices on meaning or tone.
Target 8	Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.
Target 9	Identify central ideas, key events, or procedures.
Target 10	Determine intended meanings of words, including academic/tier 2 words, domain-specific/tier 3 words, and words with multiple meanings, based on context, word relationships (e.g., synonyms), word structure (e.g., common Greek or Latin roots, affixes), or use of resources (e.g., dictionary, glossary), with primary focus on the academic vocabulary common to complex texts in all disciplines.
Target 11	Make an inference or provide a conclusion and use supporting evidence to justify/explain inferences (e.g., author's line of reasoning, point of view/purpose, relevance of evidence or elaboration to support claims, concepts, ideas).
Target 12	Interpret, explain, or connect information presented within or across texts (e.g., compare/contrast, cause/effect, integrate information).
Target 13	Relate knowledge of text structures or text features (e.g., graphs, charts, timelines) to obtain, interpret, explain, or integrate information.

Figure B.6: Content Emphases for ELA Page 6

Target 14	Interpret figurative language, literary devices, or connotative meanings of words and phrases used in context and the impact of those word choices on meaning or tone.

Grade 4, Claim 2 Targets

Target 1a	Demonstrate ability to use specific narrative techniques (e.g., dialogue sensory or concrete details, description), chronology, appropriate transitional strategies for coherence, or authors' craft appropriate to purpose (e.g., closure, detailing characters, plot, setting, an event).
Target 1b	Revise one or more paragraphs demonstrating specific narrative techniques (e.g., dialogue, sensory or concrete details, description), chronology, appropriate transitional strategies for coherence, or authors' craft appropriate to purpose (e.g., closure, detailing characters, plot, setting, an event).
Target 3a	Demonstrate ability to organize ideas in informational/explanatory texts by stating a focus (main idea), including appropriate transitional strategies for coherence, or supporting evidence and elaboration, or writing body paragraphs, or a conclusion that is appropriate to purpose and audience and related to the information or explanation presented.
Target 3b	Revise one or more informational/explanatory paragraphs demonstrating ability to organize ideas by stating a focus (main idea), including appropriate transitional strategies for coherence, or supporting evidence and elaboration, or writing body paragraphs, or a conclusion that is appropriate to purpose and audience and related to the information or explanation presented.
Target 6a	Demonstrate ability to state an opinion about topics or sources; set a context, organize ideas, develop supporting evidence/reasons and elaboration, or develop a conclusion that is appropriate to purpose and audience and related to the opinion presented.
Target 6b	Revise one or more paragraphs demonstrating ability to state opinions about topics or sources; set a context, organize ideas, develop supporting evidence/reasons and elaboration, or develop a conclusion appropriate to purpose and audience and related to the opinion presented.
Target 8	Strategically use language and vocabulary (including academic or domain-specific vocabulary) appropriate to the purpose and audience when revising or composing texts.
Target 9	Apply or edit grade-appropriate grammar usage, capitalization, punctuation, and spelling to clarify a message and edit narrative, explanatory/informational, and opinion texts.

Figure B.7: Content Emphases for ELA Page 7

Grade 4, Claim 3 Target

Target 4	Interpret and use information delivered orally.

Figure B.8: Content Emphases for ELA Page 8

Grade 5, Claim 1 Targets

Target 1	Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.
Target 2	Identify or summarize central ideas, key events, the sequence of events, or the author's message or purpose presented in a text.
Target 3	Determine intended or precise meanings of words, including multiple meanings of academic/tier 2 words, based on context, word relationships (e.g., antonyms, homographs), word structure (e.g., common Greek or Latin roots, affixes), or use of resources (e.g., dictionary, thesaurus), with primary focus on determining meaning based on context and the academic/tier 2 vocabulary common to complex texts in all disciplines.
Target 4	Make an inference or provide a conclusion and use supporting evidence to justify/explain inferences (e.g., character development/actions/traits; first- or third-person point of view; theme; author's message or purpose).
Target 5	Examine or compare relationships (literary elements: setting, conflict, dialogue, point of view, characterization) within or across texts.
Target 6	Analyze text structures, genre-specific features, or formats (visual/graphic/auditory effects) of texts and the impact of those choices on meaning or presentation.
Target 7	Interpret figurative language (e.g., metaphors, similes, idioms), literary devices, or connotative meanings of words and phrases used in context and the impact of those word choices on meaning or tone.
Target 8	Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.
Target 9	Identify central ideas, key events, procedures, or topics and subtopics.
Target 10	Determine intended meanings of words including academic/tier 2 words, domain-specific/tier 3 words, and words with multiple meanings, based on context, word relationships (e.g., synonyms), word structure (e.g., common Greek or Latin roots, affixes), or use of resources (e.g., dictionary, glossary), with primary focus on the academic vocabulary common to complex texts in all disciplines.
Target 11	Make an inference or provide a conclusion and use supporting evidence to justify/explain inferences (e.g., author's line of reasoning, point of view/purpose, relevance of evidence or elaboration to support claims, concepts, ideas).
Target 12	Analyze or compare how information is presented within or across texts (e.g., events, people, ideas, topic).
Target 13	Relate knowledge of text structures to obtain, interpret, explain, or integrate information or to compare or connect information across texts.

Figure B.9: Content Emphases for ELA Page 9

Target 14	Interpret figurative language (e.g., metaphors, similes, idioms), literary devices, or connotative meanings of words and phrases used in context and the impact of those word choices on meaning or tone.

Grade 5, Claim 2 Targets

Target 1a	Demonstrate ability to use specific narrative techniques (e.g., dialogue sensory or concrete details, description), chronology, appropriate transitional strategies for coherence, or authors' craft appropriate to purpose (e.g., closure, detailing characters, plot, setting, or an event).
Target 1b	Revise one or more paragraphs demonstrating specific narrative techniques (e.g., dialogue, description), chronology, appropriate transitional strategies for coherence, or authors' craft.
Target 3a	Demonstrate ability to organize ideas in informational/explanatory texts by stating a focus (main idea), including appropriate transitional strategies for coherence, or supporting evidence and elaboration, or writing body paragraphs, or a conclusion that is appropriate to purpose and audience and related to the information or explanation presented.
Target 3b	Revise one or more informational/explanatory paragraphs demonstrating ability to organize ideas by stating a focus (main idea), including appropriate transitional strategies for coherence, or supporting evidence and elaboration, or writing body paragraphs, or a conclusion that is appropriate to purpose and audience and related to the information or explanation presented.
Target 6a	Demonstrate ability to state opinions about topics or sources; set a context, organize ideas, develop supporting evidence/reasons and elaboration, or develop a conclusion that is appropriate to purpose and audience and related to the opinion presented.
Target 6b	Revise one or more paragraphs demonstrating ability to state opinions about topics or sources; set a context, organize ideas, develop supporting evidence/reasons and elaboration, or develop a conclusion appropriate to purpose and audience and related to the opinion presented.
Target 8	Strategically use language and vocabulary (including academic or domain-specific vocabulary) appropriate to the purpose and audience when revising or composing texts.
Target 9	Apply or edit grade-appropriate grammar usage, capitalization, punctuation, and spelling to clarify a message and edit narrative, explanatory/informational, and opinion texts.

Grade 5, Claim 3 Target

Target 4	Interpret and use information delivered orally.

Figure B.10: Content Emphases for ELA Page 10

Content Emphases for Grades 6-8

Claim (% of Test)	Focus	Target	Goal DOK	Relative Emphasis
1. Reading (60-65%)	Literary Texts	1: Key Details	2	Low
		2: Central Ideas	2	High
		3: Word Meanings	2	Low
		4: Reasoning & Evidence	3	High
		5: Analysis Within Or Across Texts	3	Low
		6: Text Structures & Features	3	
		7: Language Use	3	
	Informational Texts	8: Key Details	2	Medium
		9: Central Ideas	2	High
		10: Word Meanings	2	Medium
		11: Reasoning & Evidence	3	High
		12: Analysis Within Or Across Texts	3	Low
		13: Text Structures & Features	3	
		14: Language Use	3	
2. Writing (25-30%)	Write / Revise	1/3/6: Write / Revise Brief Texts	2	High
	Language / Vocabulary	8: Language & Vocabulary Use	2	High
	Conventions	9: Edit	1	High
3. Listening (10-15%)	Listen	4: Listen / Interpret	3	High

Figure B.11: Content Emphases for ELA Page 11

Grade 6, Claim 1 Targets

Target 1	Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.
Target 2	Identify or summarize central ideas, key events, the sequence of events, or the author's purpose presented in a text.
Target 3	Determine intended or precise meanings of words, including academic/tier 2 words, domain-specific/tier 3 words, and words with multiple meanings, based on context, word relationships (e.g., synonyms), word structure (e.g., common Greek or Latin roots, affixes), or use of resources (e.g., dictionary, glossary), with primary focus on determining meaning based on context and the academic/tier 2 vocabulary common to complex texts in all disciplines.
Target 4	Make an inference or provide a conclusion and use supporting evidence to justify/explain inferences (e.g., character development/actions/traits; first- or third-person point of view).
Target 5	Analyze relationships among literary elements (dialogue, advancing action, character actions/interactions, point of view) within or across texts.
Target 6	Analyze text structures, genre-specific features, or formats (visual/graphic/auditory effects) of texts and the impact of those choices on meaning or presentation.
Target 7	Interpret figurative language use (e.g., personification, metaphor), literary devices, or connotative meanings of words and phrases used in context and the impact of those word choices on meaning or tone.
Target 8	Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.
Target 9	Summarize central ideas, key events, procedures, or topics and subtopics.
Target 10	Determine intended or precise meanings of words, including domain-specific/tier 3 words and words with multiple meanings (academic/tier 2 words), based on context, word relationships (e.g., antonyms, homographs), word structure (e.g., common Greek or Latin roots, affixes), or use of resources (e.g., dictionary, glossary), with primary focus on the academic vocabulary common to complex texts in all disciplines.
Target 11	Make an inference or provide a conclusion and use supporting evidence to justify/explain inferences (e.g., author's line of reasoning, point of view/purpose, relevance of evidence or elaboration to support claims, concepts, ideas).
Target 12	Analyze or compare how information is presented within or across texts (e.g., events, people, ideas, topics) or how conflicting information across texts reveals author's point of view.

Figure B.12: Content Emphases for ELA Page 12

Target 13	Relate knowledge of text structures or genre-specific features to analyze or integrate information.
Target 14	Interpret figurative language (e.g., hyperbole, personification, analogies), use of literary devices, or connotative meanings of words and phrases used in context and the impact of those word choices on meaning or tone.

Grade 6, Claim 2 Targets

Target 1a	Demonstrate ability to use specific narrative techniques (e.g., dialogue, description) and appropriate text structures and transitional strategies for coherence in narrative text (e.g., closure, introduce narrator, use dialogue when describing an event).
Target 1b	Apply narrative techniques (e.g., dialogue, description) and appropriate text structures and transitional strategies for coherence when revising one or more paragraphs of narrative text (e.g., closure, introduce narrator, use dialogue when describing an event).
Target 3a	Demonstrate ability to apply a variety of strategies in informational/explanatory text: organizing ideas by stating and maintaining a focus (thesis)/tone, providing appropriate transitional strategies for coherence, developing a topic including relevant supporting evidence/vocabulary and elaboration, or providing a conclusion that is appropriate to purpose and audience and follows from the information or explanation presented.
Target 3b	Apply a variety of strategies when revising one or more paragraphs of informational/explanatory text: organizing ideas by stating and maintaining a focus (thesis)/tone, providing appropriate transitional strategies for coherence, developing a topic including relevant supporting evidence/vocabulary and elaboration, or providing a conclusion that is appropriate to purpose and audience and follows from the information or explanation presented.
Target 6a	Demonstrate ability to apply a variety of strategies in texts that express arguments about topics or sources: establishing and supporting a claim, organizing and citing supporting evidence using credible sources, providing appropriate transitional strategies for coherence, appropriate vocabulary, or providing a conclusion that is appropriate to purpose and audience and follows from the argument(s) presented.
Target 6b	Apply a variety of strategies when revising one or more paragraphs of text that express arguments about topics or sources: establishing and supporting a claim, organizing and citing supporting evidence using credible sources, providing appropriate transitional strategies for coherence, appropriate vocabulary, or providing a conclusion that is appropriate to purpose and audience and follows from the argument(s) presented.

Figure B.13: Content Emphases for ELA Page 13

Target 8	Strategically use precise language and vocabulary (including academic words, domain-specific vocabulary, and figurative language) and style appropriate to the purpose and audience when revising or composing texts.
Target 9	Apply or edit grade-appropriate grammar usage, capitalization, punctuation, and spelling to clarify a message and edit narrative, explanatory/informational, and argumentative texts.

Grade 6, Claim 3 Target

Target 4	Analyze, interpret, and use information delivered orally.

Figure B.14: Content Emphases for ELA Page 14

Grade 7, Claim 1 Targets

Target 1	Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.
Target 2	Identify or summarize central ideas, key events, the sequence of events, or the author's purpose presented in a text.
Target 3	Determine intended or precise meanings of words, including academic/tier 2 words, domain-specific/tier 3 words, and words with multiple meanings, based on context, word relationships (e.g., synonyms), word structure (e.g., common Greek or Latin roots, affixes), or use of resources (e.g., dictionary, glossary), with primary focus on determining meaning based on context and the academic/tier 2 vocabulary common to complex texts in all disciplines.
Target 4	Make an inference or provide a conclusion and use supporting evidence to justify/explain inferences (e.g., character development/actions/traits; first- or third-person point of view).
Target 5	Analyze relationships among literary elements (dialogue, advancing action, character actions/interactions, point of view) within or across texts.
Target 6	Analyze text structures, genre-specific features, or formats (visual/graphic/auditory effects) of texts and the impact of those choices on meaning or presentation.
Target 7	Interpret figurative language use (e.g., imagery), literary devices (e.g., flashback, foreshadowing, alliteration, onomatopoeia), or connotative meanings of words and phrases used in context and the impact of those word choices on meaning or tone.
Target 8	Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.
Target 9	Summarize central ideas, key events, procedures, or topics and subtopics.
Target 10	Determine intended or precise meanings of words, including domain-specific/tier 3 words and words with multiple meanings (academic/tier 2 words), based on context, word relationships (e.g., antonyms, homographs), word structure (e.g., common Greek or Latin roots, affixes), or use of resources (e.g., dictionary, glossary), with primary focus on determining meaning based on context and the academic/tier 2 vocabulary common to complex texts in all disciplines.
Target 11	Make an inference or provide a conclusion and use supporting evidence to justify/explain inferences (e.g., author's line of reasoning, point of view/purpose, relevance of evidence or elaboration to support claims, concepts, ideas).

Figure B.15: Content Emphases for ELA Page 15

Target 12	Analyze or compare how information is presented within or across texts (events, people, ideas, topics) or how conflicting information across texts reveals author's point of view.
Target 13	Relate knowledge of text structures and genre-specific features to compare or analyze the impact of those choices on meaning or presentation.
Target 14	Interpret figurative language (e.g., clichés, puns, hyperbole), use of literary devices, or connotative meanings of words and phrases used in context and the impact of those word choices on meaning or tone.

Grade 7, Claim 2 Targets

Target 1a	Demonstrate ability to use specific narrative techniques (e.g., dialogue, description) and appropriate text structures and transitional strategies for coherence in narrative text (e.g., closure, introduce narrator, or use dialogue when describing an event).
Target 1b	Apply narrative techniques (e.g., dialogue, description) and appropriate text structures and transitional strategies for coherence when revising one or more paragraphs of narrative text (e.g., closure, introduce narrator, use dialogue when describing an event).
Target 3a	Demonstrate ability to apply a variety of strategies in informational/explanatory text: organizing ideas by stating and maintaining a focus (thesis)/tone, providing appropriate transitional strategies for coherence, developing a topic including relevant supporting evidence/vocabulary and elaboration, or providing a conclusion that is appropriate to purpose and audience and follows from and supports the information or explanation presented.
Target 3b	Apply a variety of strategies when revising one or more paragraphs of informational/explanatory text: organizing ideas by stating and maintaining a focus (thesis)/tone, providing appropriate transitional strategies for coherence, developing a topic including relevant supporting evidence/vocabulary and elaboration, or providing a conclusion that is appropriate to purpose and audience and related to the information or explanation presented.
Target 6a	Demonstrate ability to apply a variety of strategies in texts that express arguments about topics or sources: establishing and supporting a claim, organizing and citing supporting evidence using credible sources, providing appropriate transitional strategies for coherence, appropriate vocabulary, or providing a conclusion that is appropriate to purpose and audience and follows from and supports the argument(s) presented.

Figure B.16: Content Emphases for ELA Page 16

Target 6b	Apply a variety of strategies when revising one or more paragraphs of text that express arguments about topics or sources: establishing and supporting a claim, organizing and citing supporting evidence using credible sources, providing appropriate transitional strategies for coherence, appropriate vocabulary, or providing a conclusion that is appropriate to purpose and audience and follows from and supports the argument(s) presented.
Target 8	Strategically use precise language and vocabulary (including academic words, domain-specific vocabulary, and figurative language) and style appropriate to the purpose and audience when revising or composing texts.
Target 9	Apply or edit grade-appropriate grammar usage, capitalization, punctuation, and spelling to clarify a message and edit narrative, explanatory/informational, and argumentative texts.

Grade 7, Claim 3 Target

Target 4	Analyze, interpret, and use information delivered orally.

Figure B.17: Content Emphases for ELA Page 17

Grade 8, Claim 1 Targets

Target 1	Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.
Target 2	Identify or summarize central ideas, key events, the sequence of events, or the author's purpose presented in a text.
Target 3	Determine intended or precise meanings of words, including academic/tier 2 words, domain-specific/tier 3 words, and words with multiple meanings, based on context, word relationships (e.g., synonyms), word structure (e.g., common Greek or Latin roots, affixes), or use of resources (e.g., dictionary, glossary), with primary focus on the academic vocabulary common to complex texts in all disciplines.
Target 4	Make an inference or provide a conclusion and use supporting evidence to justify/explain inferences (e.g., character development/actions/traits; first- or third-person point of view).
Target 5	Analyze relationships among literary elements (e.g., dialogue, advancing action, character actions/interactions, point of view) within or across texts.
Target 6	Analyze text structures, genre-specific features, or formats (visual/graphic/auditory effects) of texts and the impact of those choices on meaning or presentation.
Target 7	Interpret figurative language, literary devices, or connotative meanings of words and phrases used in context and the impact of those word choices on meaning or tone.
Target 8	Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.
Target 9	Summarize central ideas, topics/subtopics, key events, or procedures using supporting ideas and details.
Target 10	Determine intended or precise meanings of words, including domain-specific/tier 3 words and words with multiple meanings (academic/tier 2 words), based on context, word relationships (e.g., antonyms, homographs), word structure (e.g., common Greek or Latin roots, affixes), or use of resources (e.g., dictionary, glossary), with primary focus on determining meaning based on context and the academic/tier 2 vocabulary common to complex texts in all disciplines.
Target 11	Make an inference or provide a conclusion and use supporting evidence to justify/explain inferences (e.g., author's line of reasoning, point of view/purpose, relevance of evidence or elaboration to support claims, concepts, ideas).
Target 12	Analyze or compare how information is presented within or across texts (e.g., events, people, ideas, topics) or how conflicting information across texts reveals author's point of view.

Figure B.18: Content Emphases for ELA Page 18

Target 13	Relate knowledge of text structures, formats, or genre-specific features (visual/graphic elements) to analyze the impact (advantages/disadvantages) on meaning or presentation.
Target 14	Interpret figurative language, literary devices, or connotative meanings of words and phrases used in context and the impact of those word choices on meaning or tone.

Grade 8, Claim 2 Targets

Target 1a	Demonstrate ability to use specific narrative techniques (e.g., dialogue, description, pacing) and appropriate text structures and transitional strategies for coherence in narrative text (e.g., closure, introduce narrator, or using dialogue when describing an event).
Target 1b	Apply narrative techniques (e.g., dialogue, description, pacing) and appropriate text structures and transitional strategies for coherence when revising one or more paragraphs of narrative text (e.g., closure, introduce narrator, or using dialogue when describing an event).
Target 3a	Demonstrate ability to apply a variety of strategies in informational/explanatory text: organizing ideas by stating and maintaining a focus (thesis) tone, providing appropriate transitional strategies for coherence, developing a topic including relevant supporting evidence/vocabulary and elaboration, or providing a conclusion that is appropriate to purpose and audience and follows from and supports the information or explanation presented.
Target 3b	Apply a variety of strategies when revising one or more paragraphs of informational/explanatory text: organizing ideas by stating and maintaining a focus (thesis)/tone, providing appropriate transitional strategies for coherence, developing a topic including relevant supporting evidence/vocabulary and elaboration, or providing a conclusion that is appropriate to purpose and audience and related to the information or explanation presented.
Target 6a	Demonstrate ability to apply a variety of strategies in texts that express arguments about topics or sources: establishing and supporting a claim, organizing and citing supporting evidence using credible sources, providing appropriate transitional strategies for coherence, appropriate vocabulary, or providing a conclusion that is appropriate to purpose and audience and follows from and supports the argument(s) presented.
Target 6b	Apply a variety of strategies when revising one or more paragraphs of text that express arguments about topics or texts: establishing and supporting a claim, organizing and citing supporting evidence using credible sources, providing appropriate transitional strategies for coherence, appropriate vocabulary, or providing a conclusion that is appropriate to purpose and audience and follows from and supports the argument(s) presented.

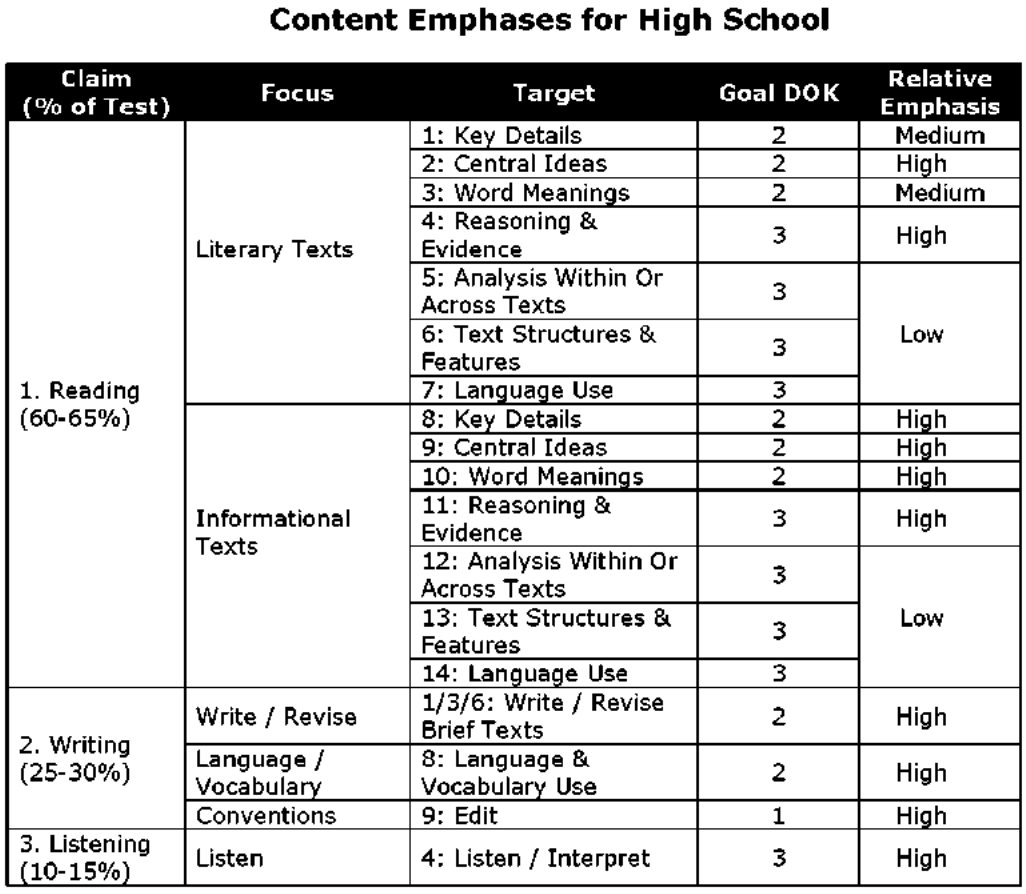
Figure B.19: Content Emphases for ELA Page 19

Target 8	Strategically use precise language and vocabulary (including academic words, domain-specific vocabulary, and figurative language) and style appropriate to the purpose and audience when revising or composing texts.
Target 9	Apply or edit grade-appropriate grammar usage, capitalization, punctuation, and spelling to clarify a message and edit narrative, explanatory/informational, and argumentative texts.

Grade 8, Claim 3 Target

Target 4	Analyze, interpret, and use information delivered orally.

Figure B.20: Content Emphases for ELA Page 20



Content Emphases for High School

Claim (% of Test)	Focus	Target	Goal DOK	Relative Emphasis
1. Reading (60-65%)	Literary Texts	1: Key Details	2	Medium
		2: Central Ideas	2	High
		3: Word Meanings	2	Medium
		4: Reasoning & Evidence	3	High
		5: Analysis Within Or Across Texts	3	Low
		6: Text Structures & Features	3	
		7: Language Use	3	
	Informational Texts	8: Key Details	2	High
		9: Central Ideas	2	High
		10: Word Meanings	2	High
		11: Reasoning & Evidence	3	High
		12: Analysis Within Or Across Texts	3	Low
		13: Text Structures & Features	3	
		14: Language Use	3	
2. Writing (25-30%)	Write / Revise	1/3/6: Write / Revise Brief Texts	2	High
	Language / Vocabulary	8: Language & Vocabulary Use	2	High
	Conventions	9: Edit	1	High
3. Listening (10-15%)	Listen	4: Listen / Interpret	3	High

Figure B.21: Content Emphases for ELA Page 21

Grade 10, Claim 1 Targets

Target 1	Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.
Target 2	Identify or summarize central ideas, key events, or the sequence of events presented in a text.
Target 3	Determine intended, precise, or nuanced meanings of words, including distinguishing connotation/denotation, analogies, and words with multiple meanings of academic/tier 2 words, based on context, word patterns, word relationships, etymology, dialectical English, idiomatic expressions, or use of specialized resources (e.g., dictionary, thesaurus), with primary focus on determining meaning based on context and the academic/tier 2 vocabulary common to complex texts in all disciplines.
Target 4	Make an inference or provide a conclusion and use supporting evidence to justify/explain inferences (e.g., character development/actions/traits; first- or third-person point of view).
Target 5	Analyze interrelationships among literary elements (e.g., characterization, conflict, ordering of actions, setting, dialogue, point of view) within or across texts.
Target 6	Analyze text structures, genre-specific features, or formats (visual/graphic/auditory effects) of texts and the impact of those choices on meaning or presentation.
Target 7	Interpret or analyze the figurative or connotative meanings of words and phrases used in context and the impact of those word choices on meaning and tone.
Target 8	Given an inference or conclusion, use explicit details and implicit information from the text to support the inference or conclusion provided.
Target 9	Summarize central ideas, topics/subtopics, key events, or procedures using supporting ideas and relevant details.
Target 10	Determine intended or precise meanings of words, including academic/tier 2 words, domain-specific/technical/tier 3 words, analogies, and connotation/denotation, based on context, word patterns, relationships, etymology, dialectical English, idiomatic expressions, or use of specialized resources (e.g., dictionary, glossary), with primary focus on the academic vocabulary common to complex texts in all disciplines.
Target 11	Make an inference or provide a conclusion and use supporting evidence to justify/explain inferences (e.g., author's line of reasoning, point of view/purpose, relevance of evidence or elaboration to support claims, concepts, ideas).

Figure B.22: Content Emphases for ELA Page 22

Target 12	Analyze texts to determine how connections are made in development of complex ideas or events or in development of topics, or rhetorical features.
Target 13	Relate knowledge of text structures or formats, or genre features (e.g., graphic/visual information), to integrate information or analyze the impact on meaning or presentation.
Target 14	Analyze the figurative or connotative meanings of words and phrases used in context and the impact of these word choices on meaning and tone.

Grade 10, Claim 2 Targets

Target 1a	Demonstrate ability to use specific narrative techniques (e.g., dialogue, description, pacing) and appropriate text structures and transitional strategies for coherence when writing one or more paragraphs of narrative text (e.g., closure, introduce narrator, or using dialogue when describing an event).
Target 1b	Apply narrative techniques (e.g., dialogue, description, pacing) and appropriate text structures and transitional strategies for coherence when revising one or more paragraphs of narrative text (e.g., closure, introduce narrator's point of view, or using dialogue when describing an event or to advance action).
Target 3a	Demonstrate ability to apply a variety of strategies in informational/explanatory text: organizing ideas by stating and maintaining a focus/tone; providing appropriate transitional strategies for coherence; developing a complex topic and subtopics, including relevant supporting evidence/vocabulary and elaboration; or providing a conclusion that is appropriate to purpose and audience and follows from and supports the information or explanation presented (e.g., articulating implications or the significance of a topic).
Target 3b	Apply a variety of strategies when writing one or more paragraphs of informational texts: organizing ideas by stating a thesis and maintaining a focus, developing a complex topic/subtopics, including relevant supporting evidence (from texts when appropriate) and elaboration, or providing a conclusion that is appropriate to purpose and audience and follows from and supports the information or explanation presented (i.e., articulating implications or the significance of a topic).
Target 6a	Demonstrate ability to apply a variety of strategies in texts that express arguments about topics or sources: establishing and supporting a precise claim, organizing and citing supporting evidence and counterclaims using credible sources, providing appropriate transitional strategies for coherence, using appropriate vocabulary, or providing a conclusion that is appropriate to purpose and audience and follows from and supports the argument(s) presented.

Figure B.23: Content Emphases for ELA Page 23

Target 6b	Apply a variety of strategies when revising one or more paragraphs of text that express arguments about topics or sources: establishing and supporting a precise claim, organizing and citing supporting evidence and counterclaims using credible sources, providing appropriate transitional strategies for coherence, using appropriate vocabulary, or providing a conclusion that is appropriate to purpose and audience and follows from and supports the argument(s) presented.
Target 8	Strategically use precise language and vocabulary (including academic and domain-specific vocabulary and figurative language) and style appropriate to the purpose and audience when revising or composing texts.
Target 9	Apply or edit grade-appropriate grammar usage, capitalization, punctuation, and spelling to clarify a message and edit narrative, explanatory/informational, and argumentative texts.

Grade 10, Claim 3 Target

Target 4	Analyze, interpret, and use information delivered orally.

End of Document

C

Item Invariance

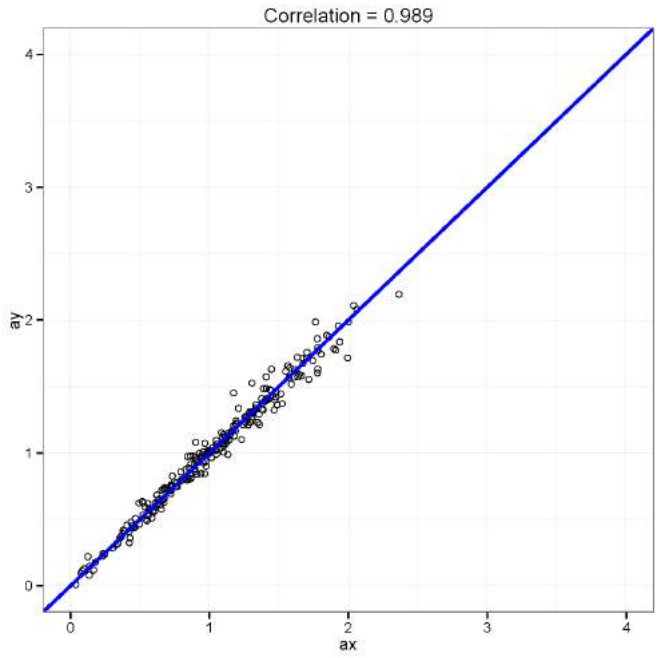


Figure C.1: Grade 3 Math Discrimination Parameter for All Items

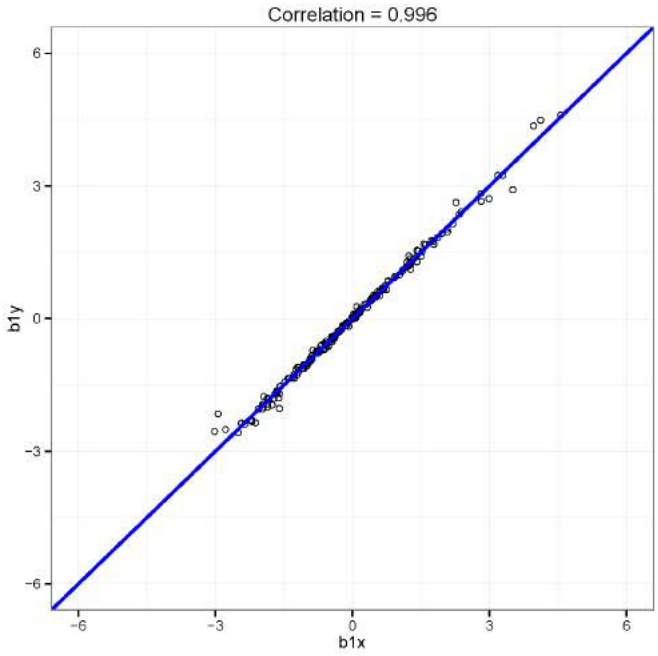


Figure C.2: Grade 3 Math Difficulty Parameter (b1) for Items with Two Score Categories

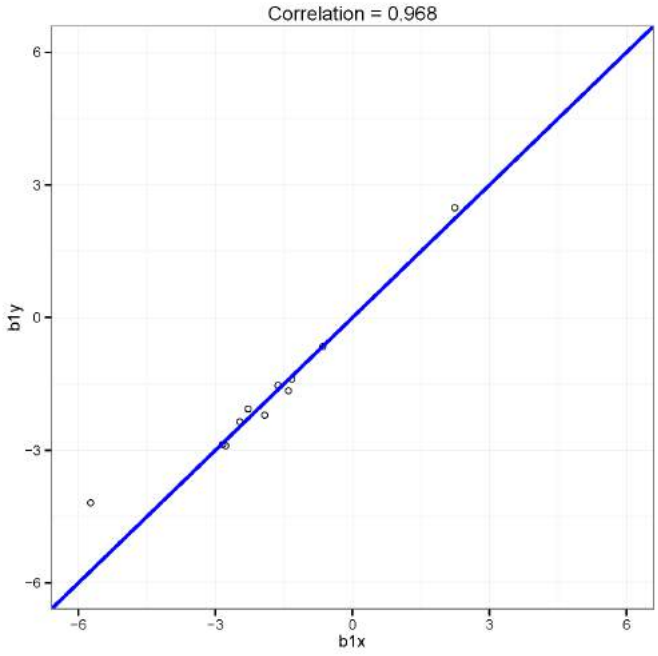


Figure C.3: Grade 3 Math Difficulty Parameter (b1) for Items with Three Score Categories

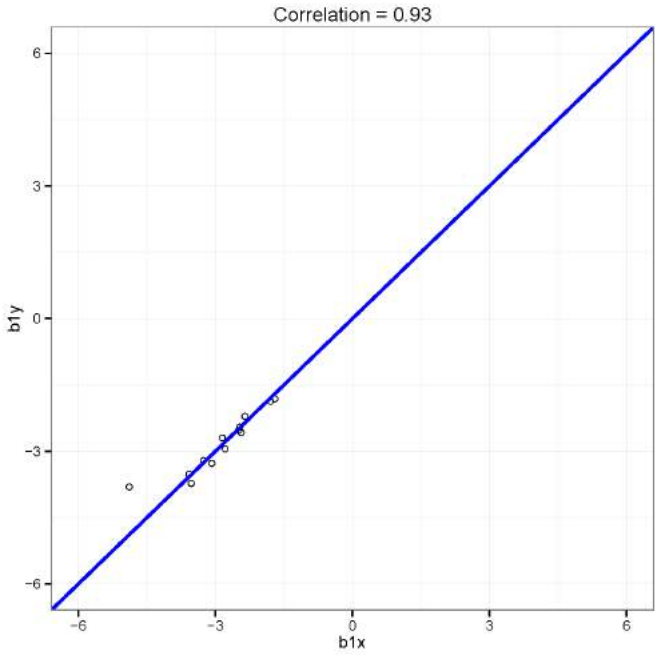


Figure C.4: Grade 3 Math Difficulty Parameter (b1) for Items with Four Score Categories

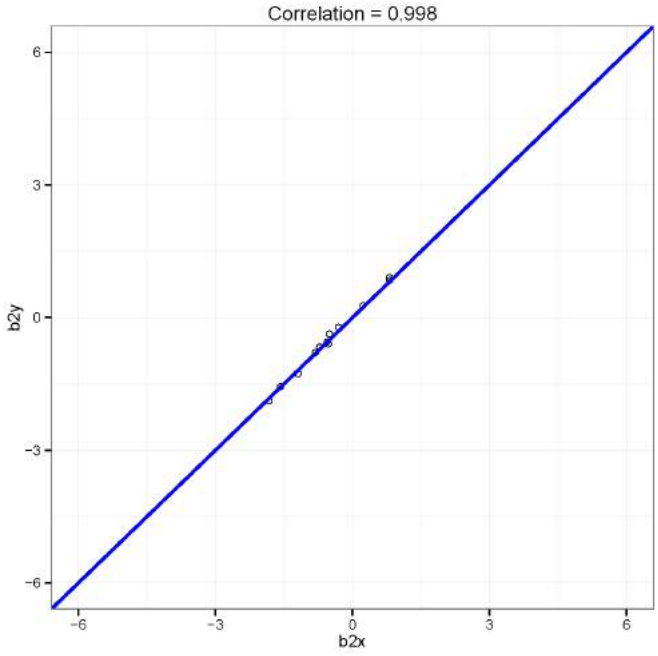


Figure C.5: Grade 3 Math Difficulty Parameter (b2) for Items with Four Score Categories

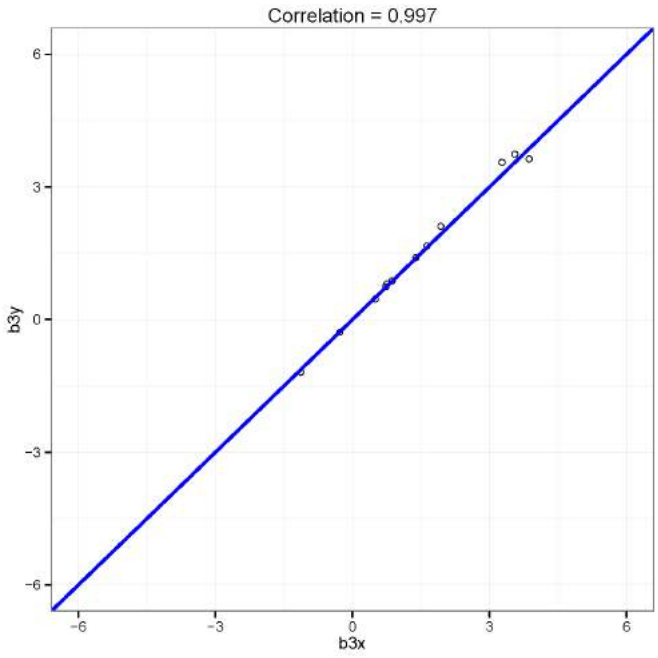


Figure C.6: Grade 3 Math Difficulty Parameter (b3) for Items with Four Score Categories

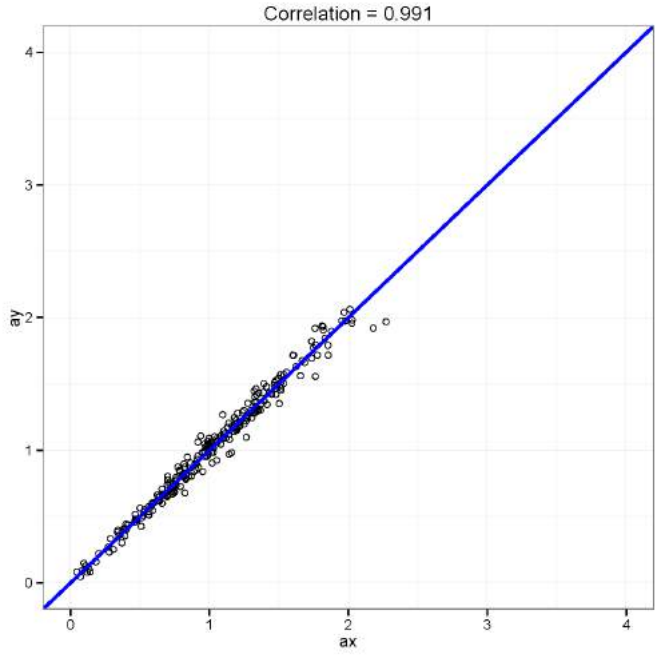


Figure C.7: Grade 4 Math Discrimination Parameter for All Items

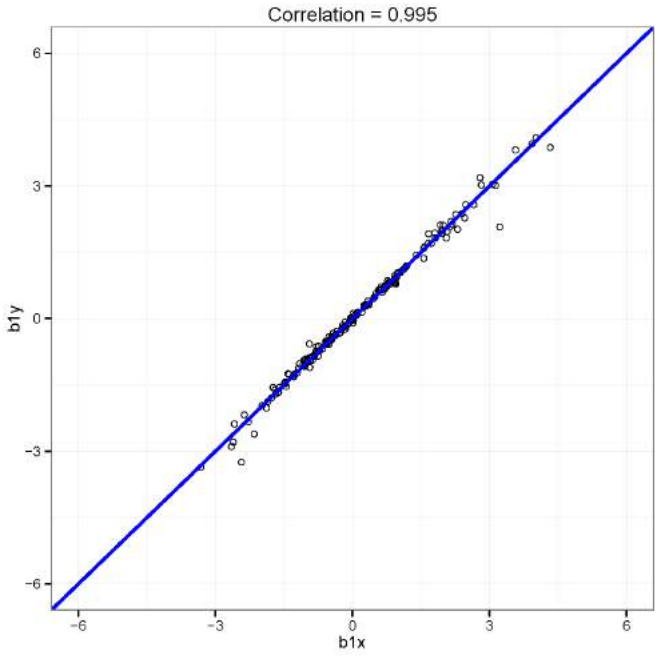


Figure C.8: Grade 4 Math Difficulty Parameter (b1) for Items with Two Score Categories

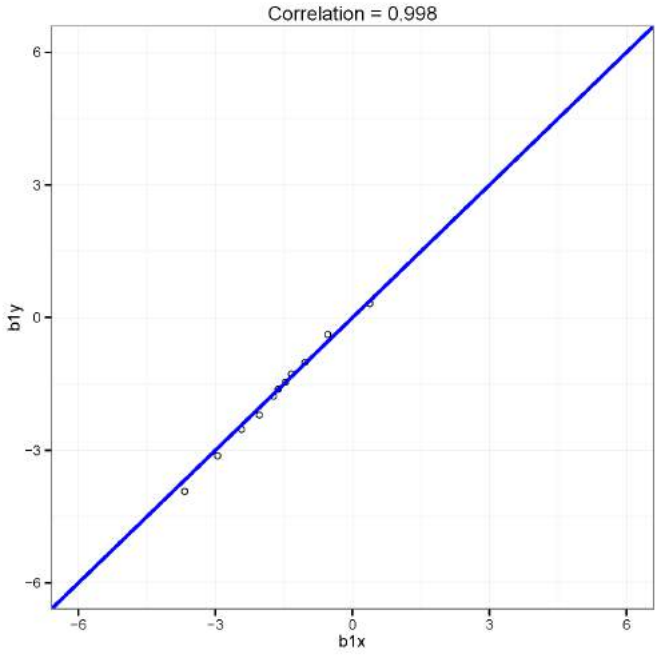


Figure C.9: Grade 4 Math Difficulty Parameter (b1) for Items with Three Score Categories

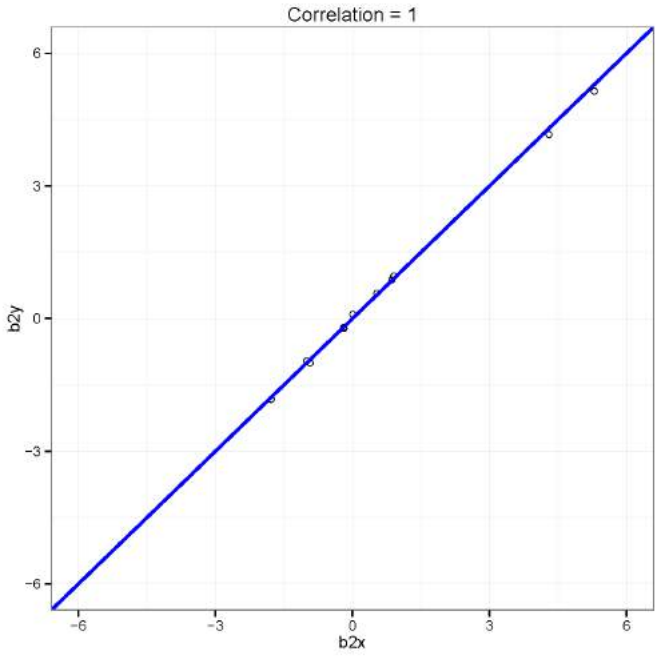


Figure C.10: Grade 4 Math Difficulty Parameter (b2) for Items with Three Score Categories

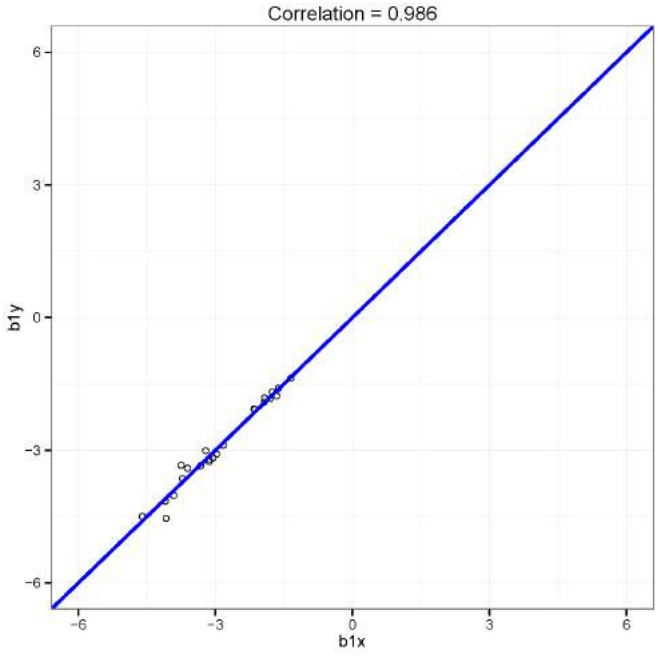


Figure C.11: Grade 4 Math Difficulty Parameter (b1) for Items with Four Score Categories

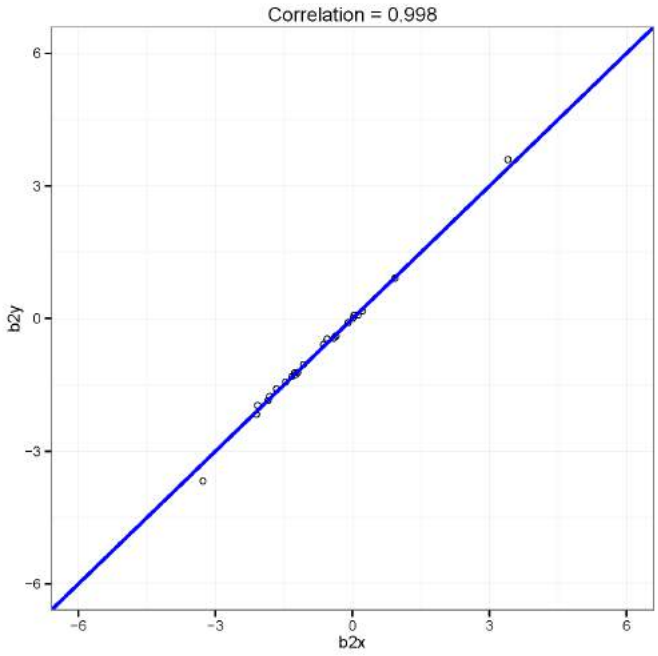


Figure C.12: Grade 4 Math Difficulty Parameter (b2) for Items with Four Score Categories

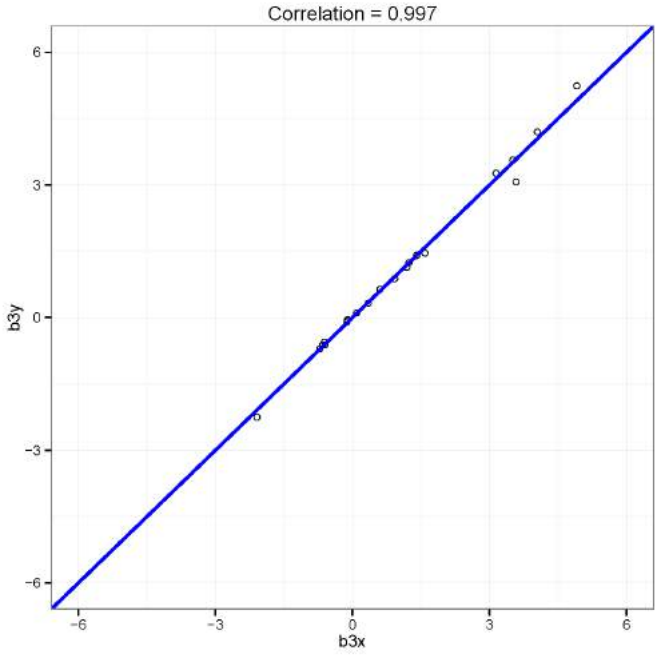


Figure C.13: Grade 4 Math Difficulty Parameter (b3) for Items with Four Score Categories

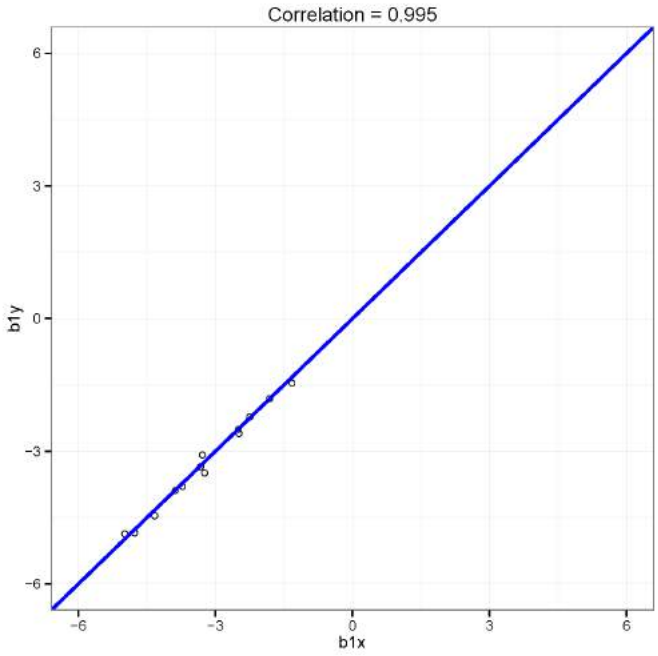


Figure C.14: Grade 4 Math Difficulty Parameter (b1) for Items with Five Score Categories

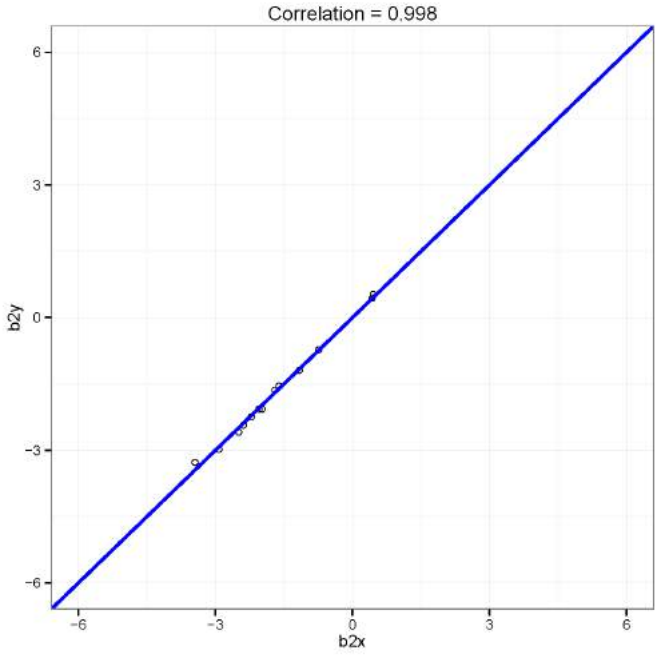


Figure C.15: Grade 4 Math Difficulty Parameter (b2) for Items with Five Score Categories

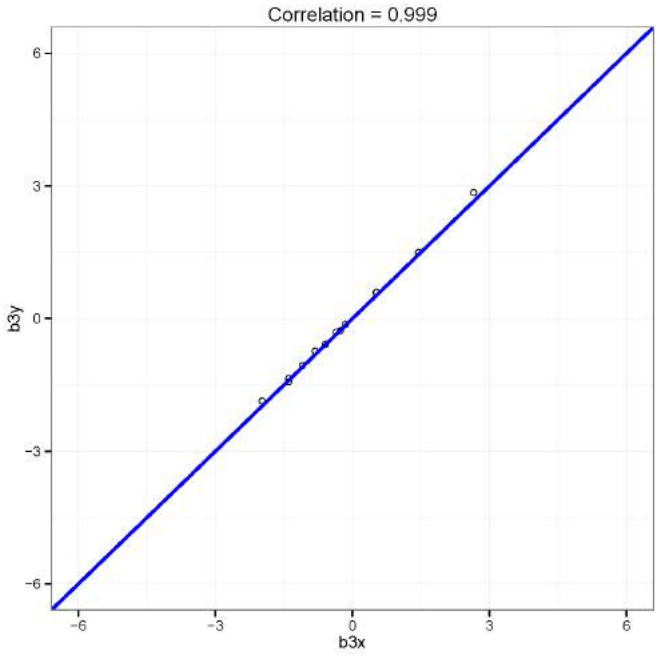


Figure C.16: Grade 4 Math Difficulty Parameter (b3) for Items with Five Score Categories

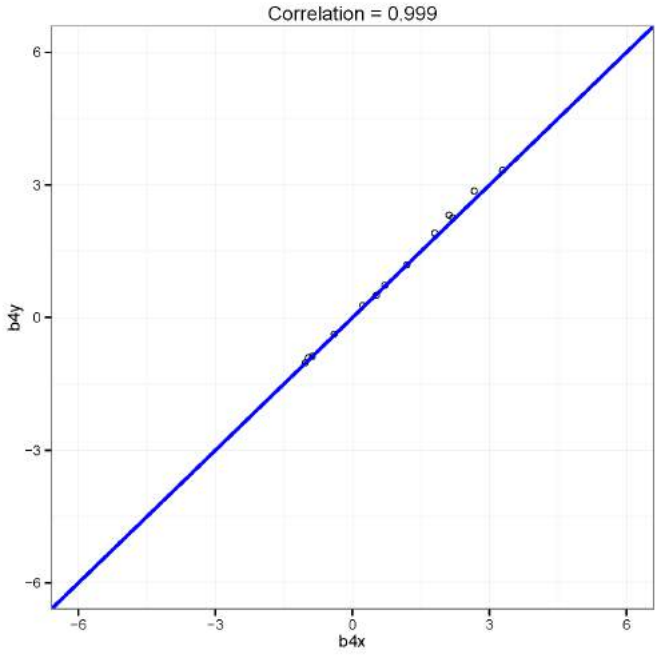


Figure C.17: Grade 4 Math Difficulty Parameter (b4) for Items with Five Score Categories

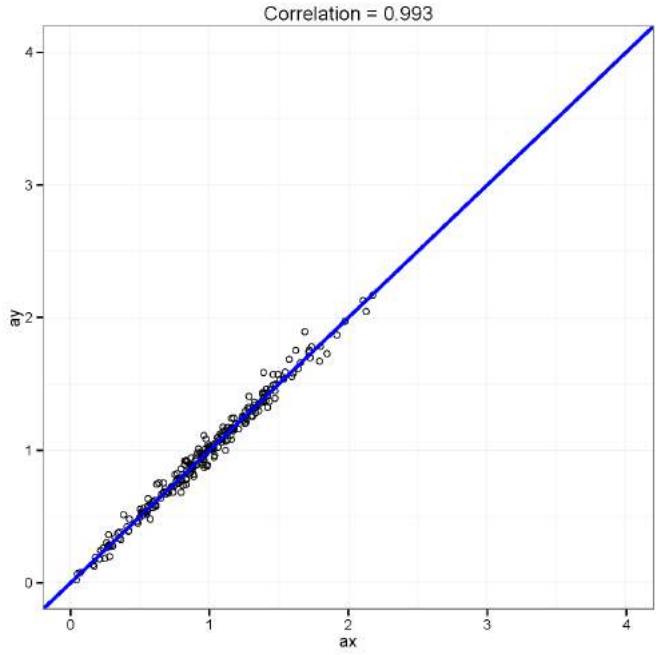


Figure C.18: Grade 5 Math Discrimination Parameter for All Items

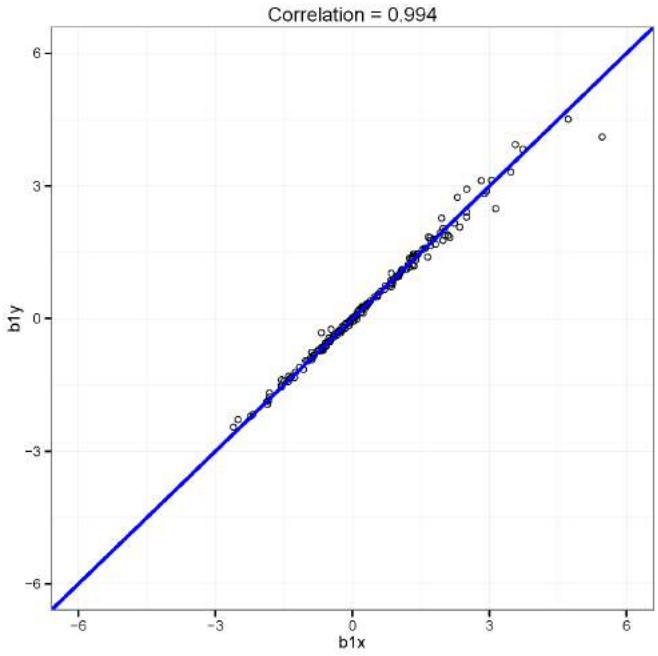


Figure C.19: Grade 5 Math Difficulty Parameter (b1) for Items with Two Score Categories

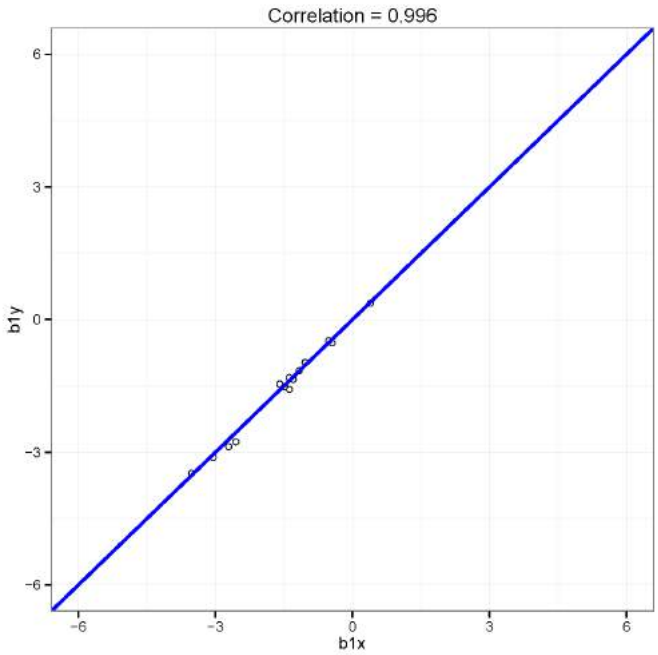


Figure C.20: Grade 5 Math Difficulty Parameter (b1) for Items with Three Score Categories

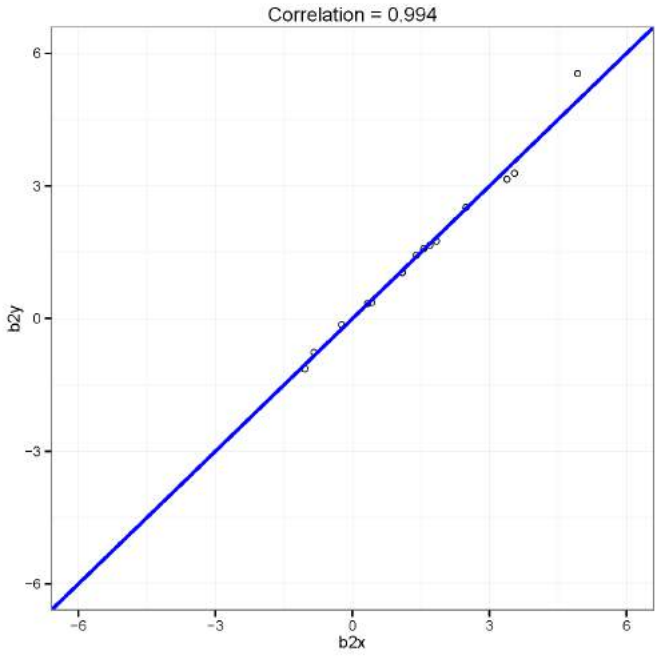


Figure C.21: Grade 5 Math Difficulty Parameter (b2) for Items with Three Score Categories

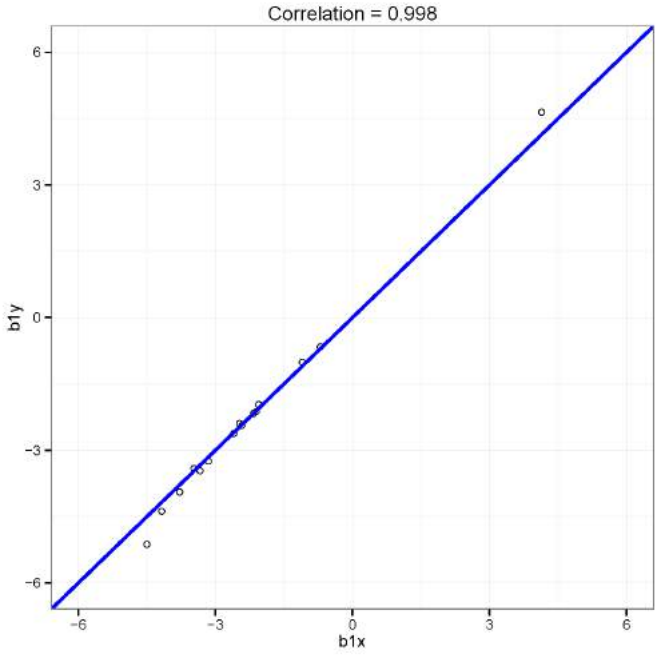


Figure C.22: Grade 5 Math Difficulty Parameter (b1) for Items with Four Score Categories

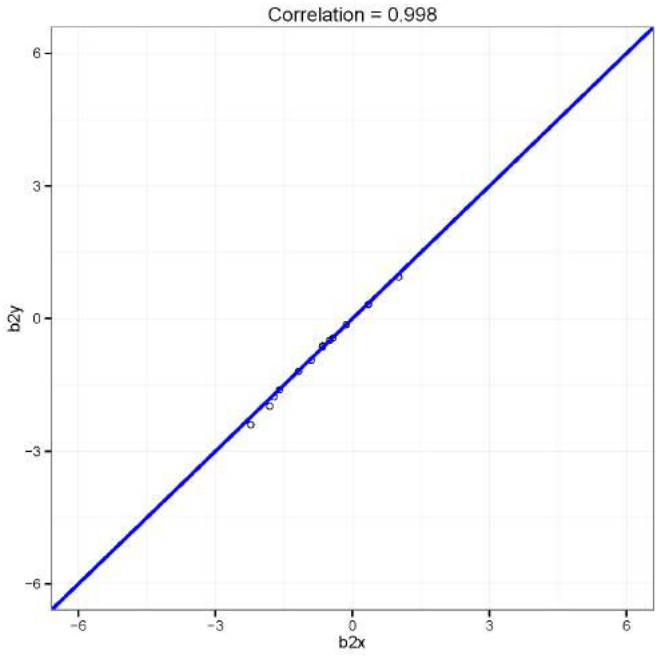


Figure C.23: Grade 5 Math Difficulty Parameter (b2) for Items with Four Score Categories

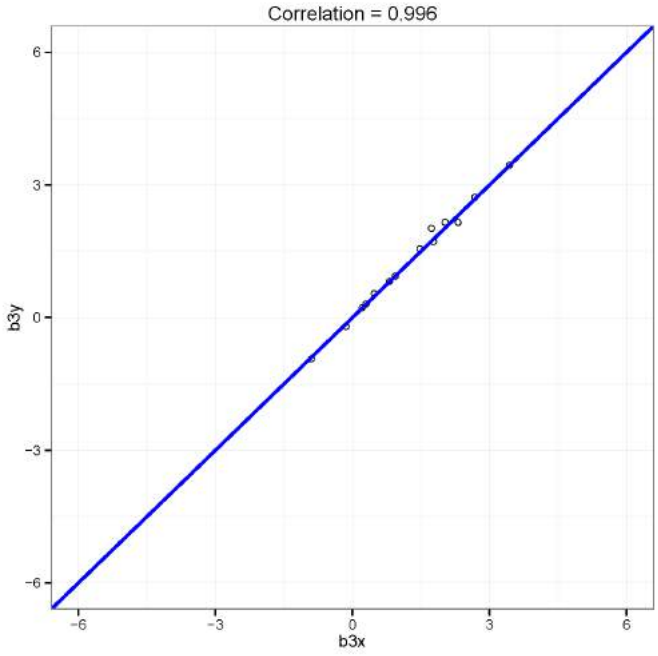


Figure C.24: Grade 5 Math Difficulty Parameter (b3) for Items with Four Score Categories

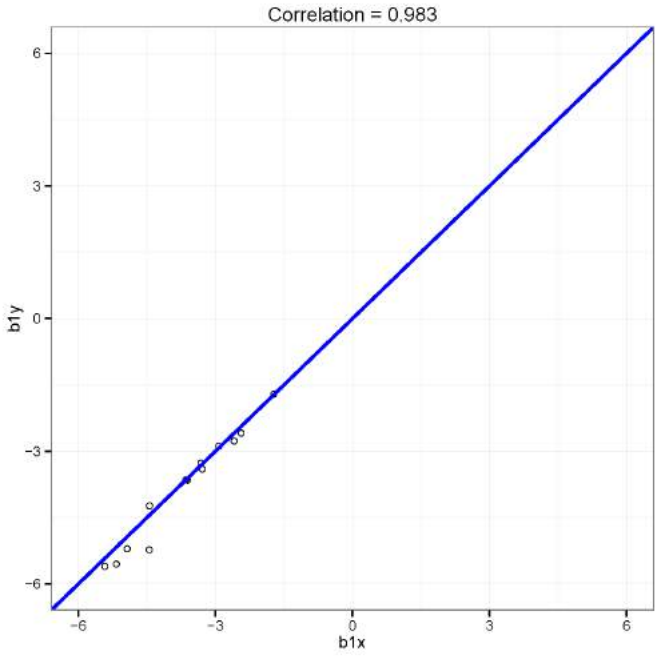


Figure C.25: Grade 5 Math Difficulty Parameter (b1) for Items with Five Score Categories

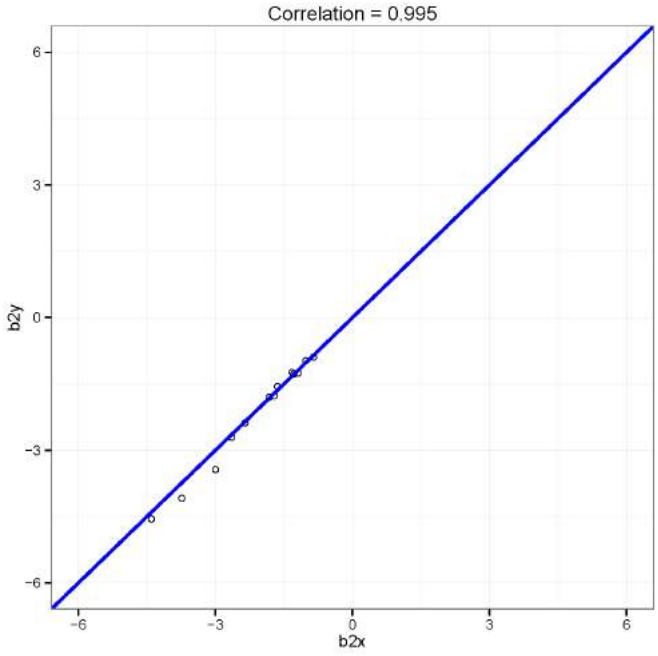


Figure C.26: Grade 5 Math Difficulty Parameter (b2) for Items with Five Score Categories

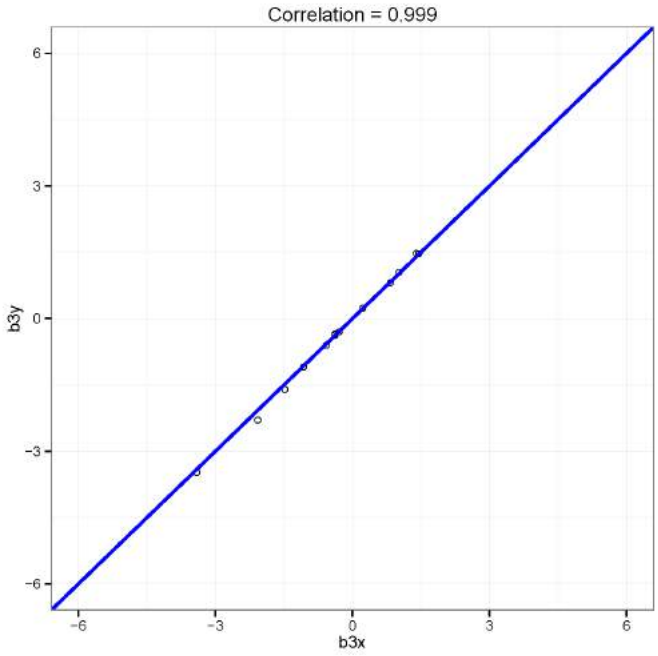


Figure C.27: Grade 5 Math Difficulty Parameter (b3) for Items with Five Score Categories

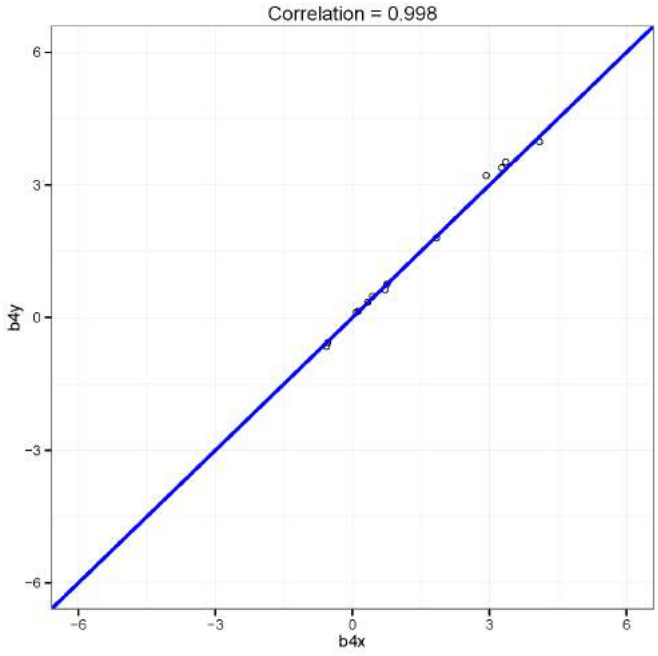


Figure C.28: Grade 5 Math Difficulty Parameter (b4) for Items with Five Score Categories

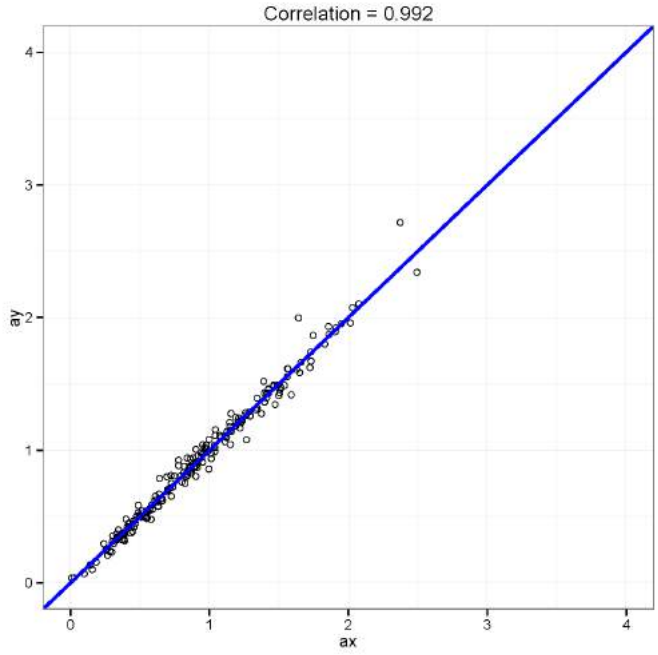


Figure C.29: Grade 6 Math Discrimination Parameter for All Items

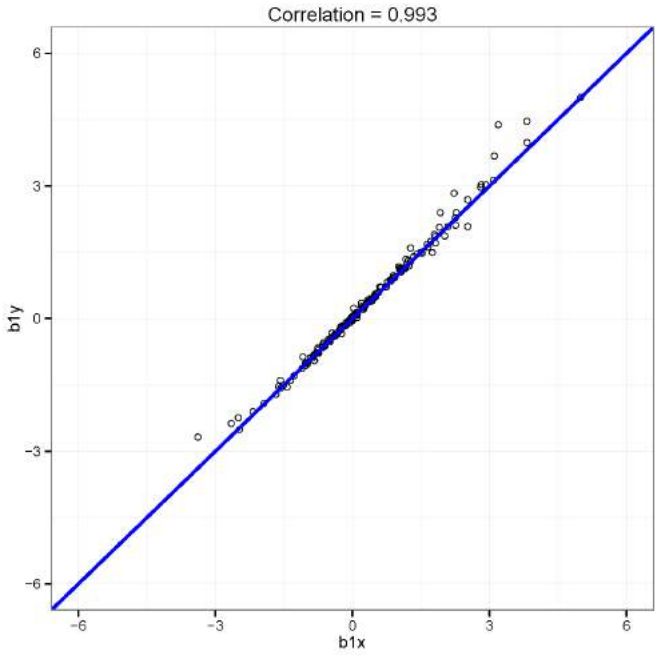


Figure C.30: Grade 6 Math Difficulty Parameter (b1) for Items with Two Score Categories

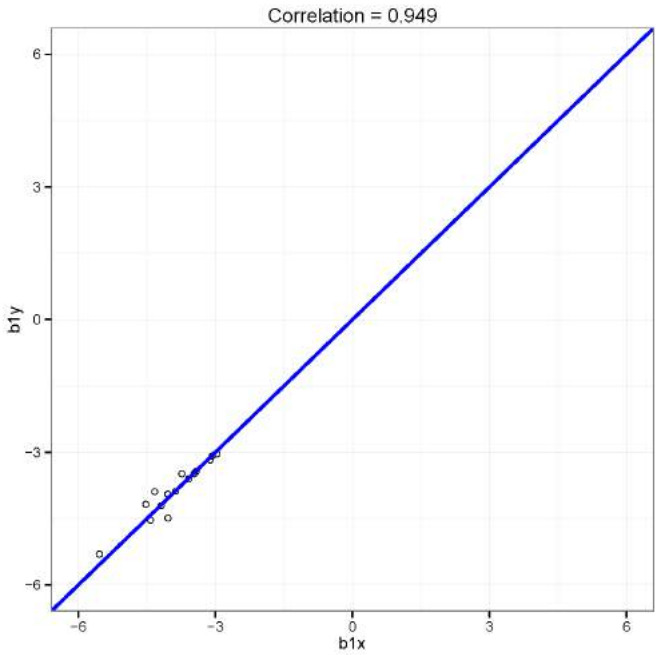


Figure C.31: Grade 6 Math Difficulty Parameter (b1) for Items with Four Score Categories

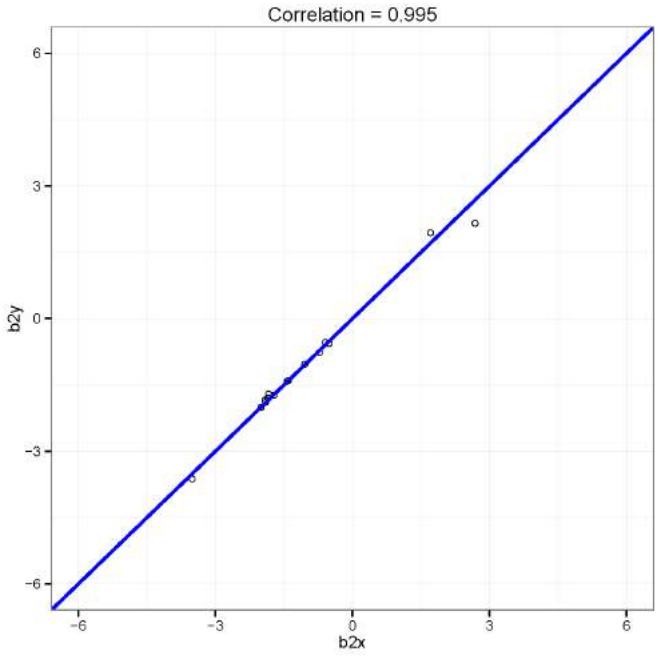


Figure C.32: Grade 6 Math Difficulty Parameter (b2) for Items with Four Score Categories

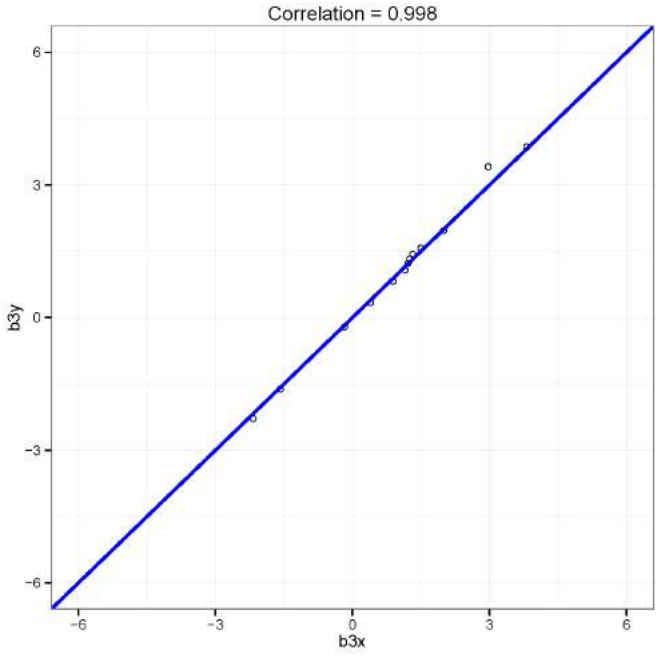


Figure C.33: Grade 6 Math Difficulty Parameter (b3) for Items with Four Score Categories

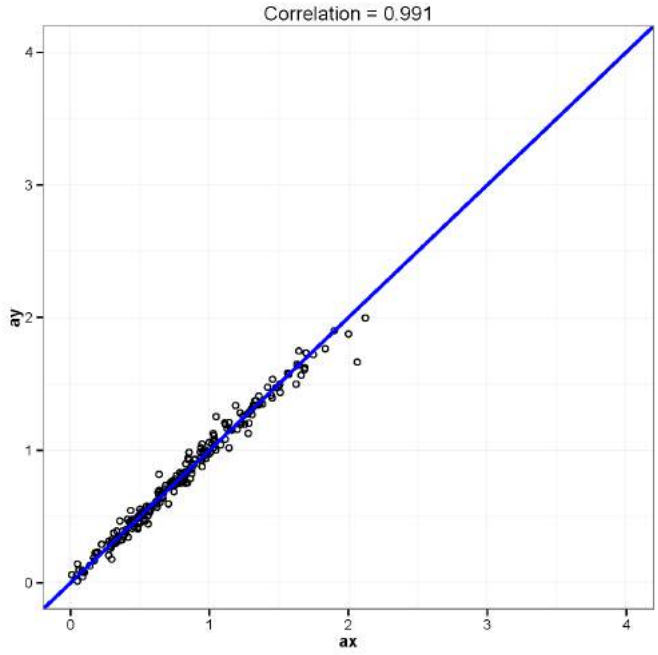


Figure C.34: Grade 7 Math Discrimination Parameter for All Items

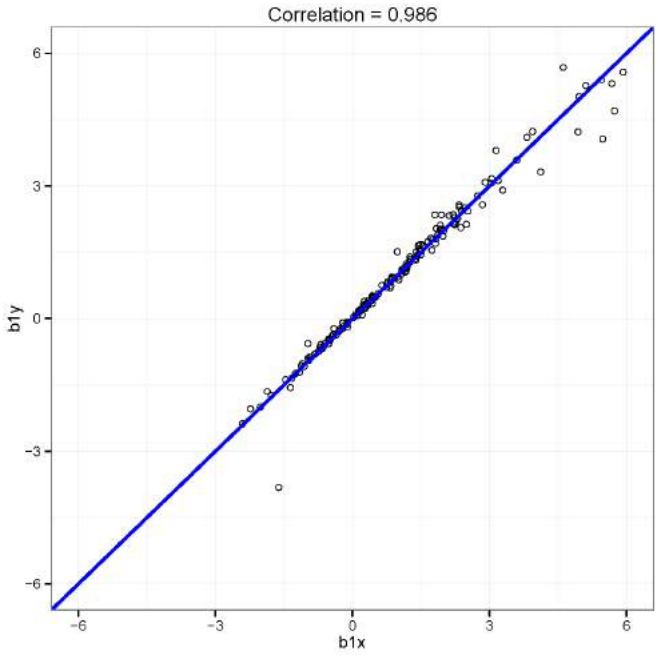


Figure C.35: Grade 7 Math Difficulty Parameter (b1) for Items with Two Score Categories

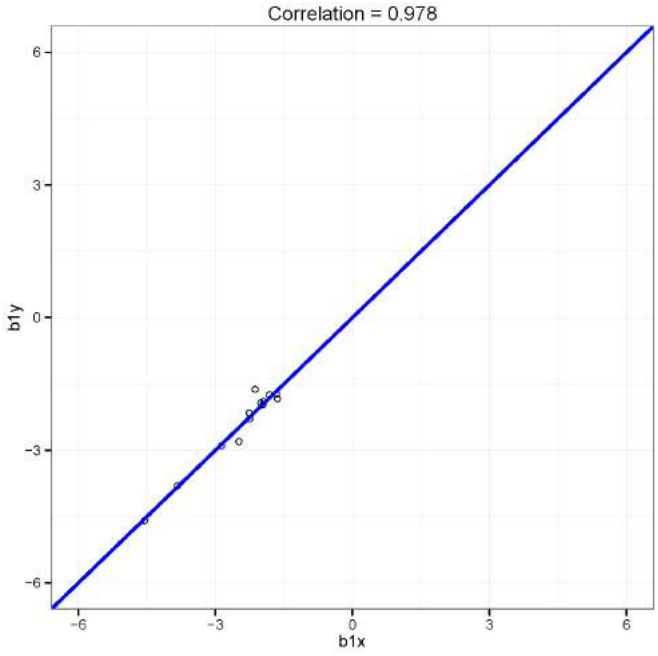


Figure C.36: Grade 7 Math Difficulty Parameter (b1) for Items with Three Score Categories

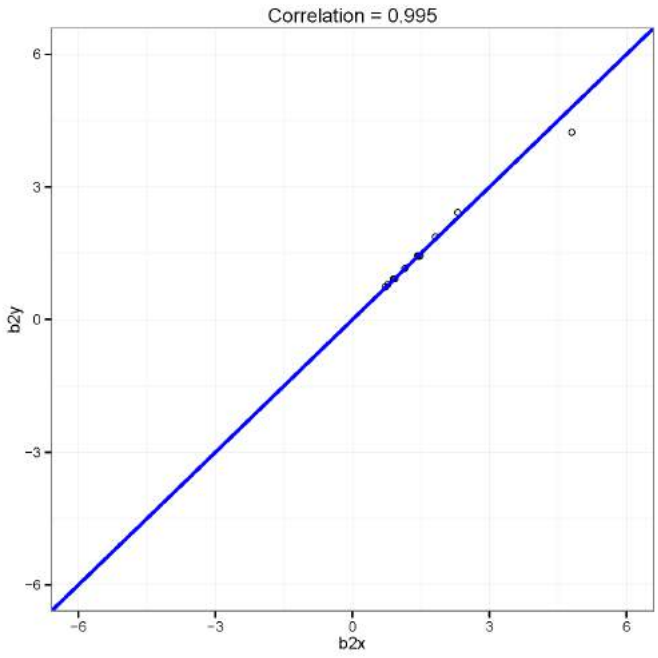


Figure C.37: Grade 7 Math Difficulty Parameter (b2) for Items with Three Score Categories

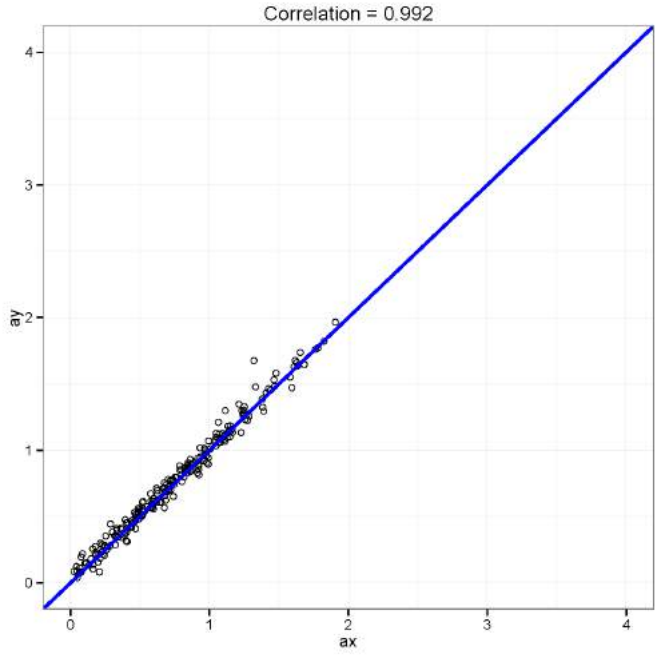


Figure C.38: Grade 8 Math Discrimination Parameter for All Items

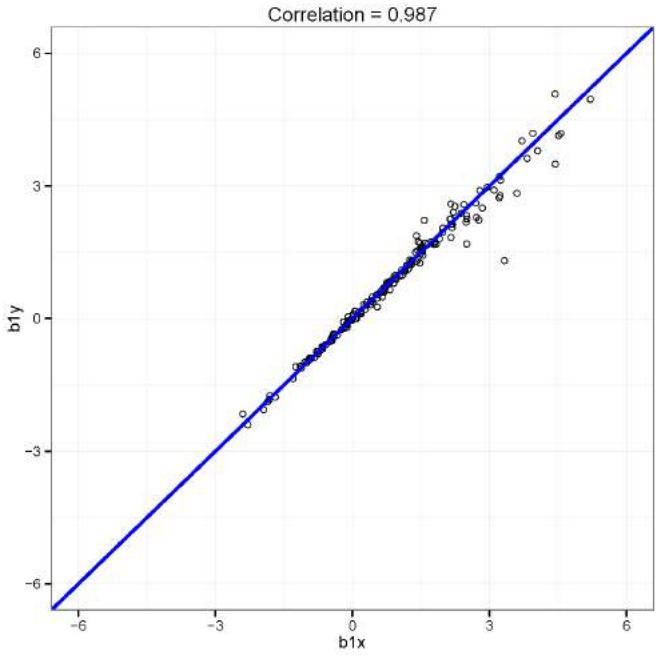


Figure C.39: Grade 8 Math Difficulty Parameter (b1) for Items with Two Score Categories

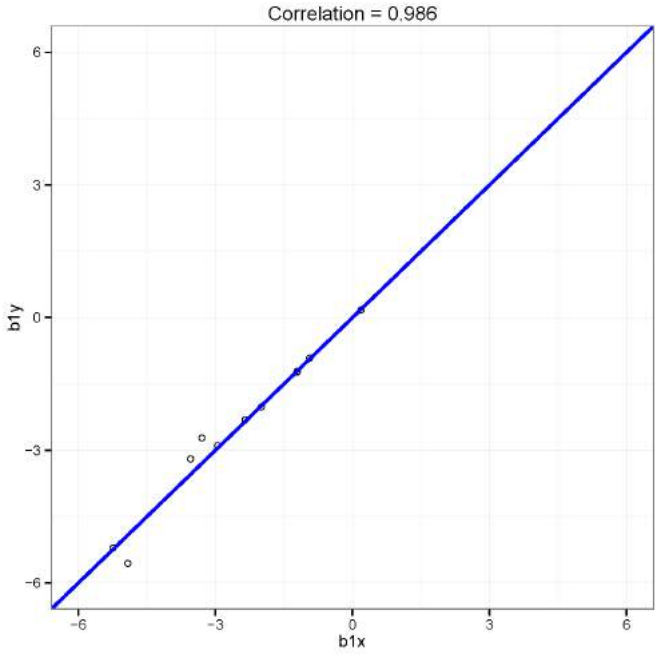


Figure C.40: Grade 8 Math Difficulty Parameter (b1) for Items with Three Score Categories

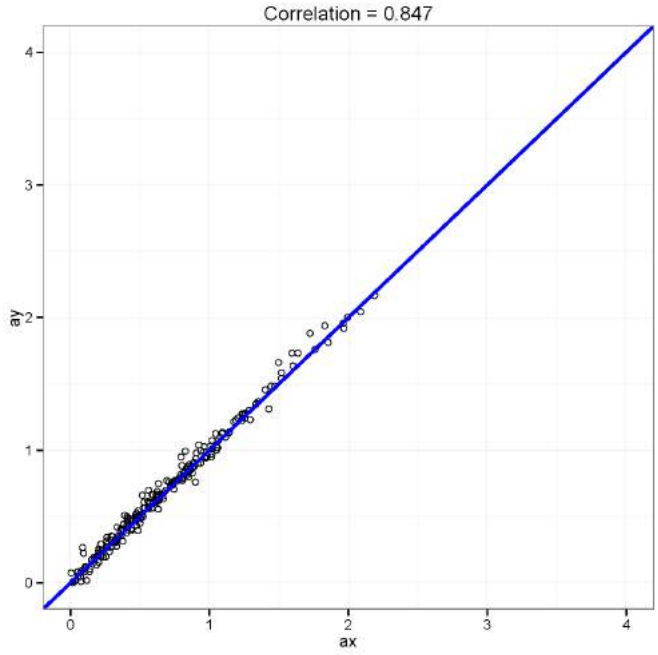


Figure C.41: Grade 10 Math Discrimination Parameter for All Items

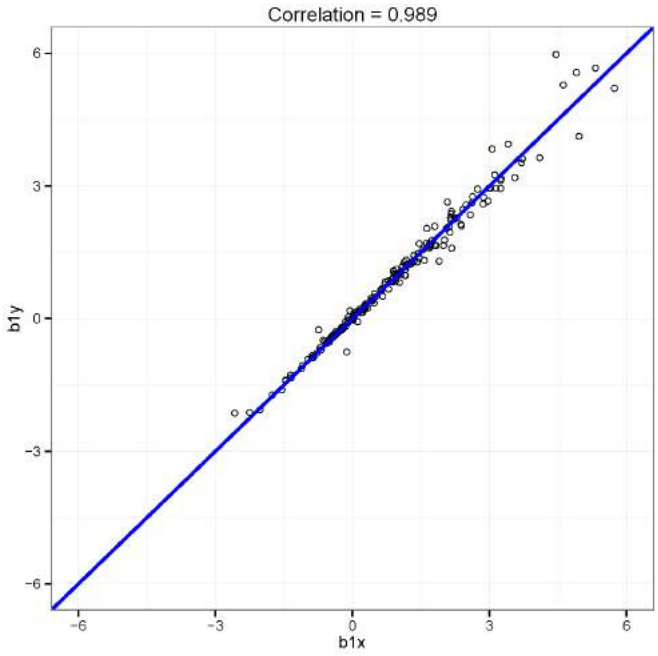


Figure C.42: Grade 10 Math Difficulty Parameter (b1) for Items with Two Score Categories

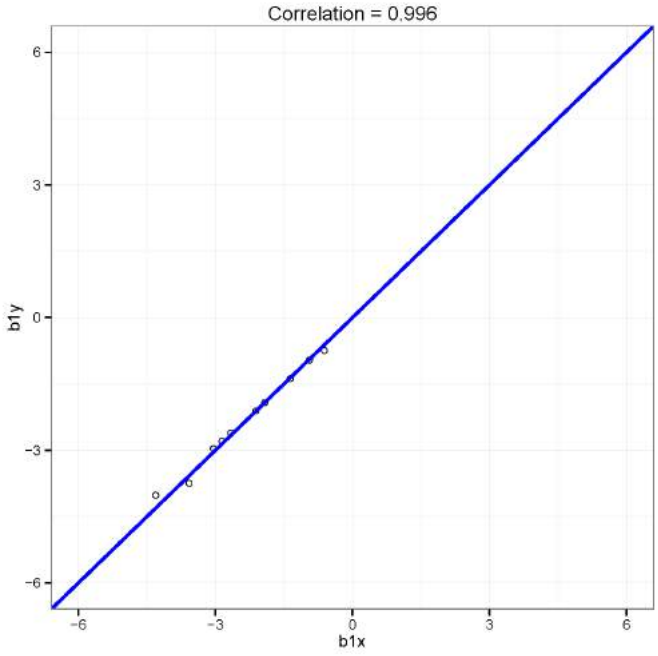


Figure C.43: Grade 10 Math Difficulty Parameter (b1) for Items with Three Score Categories

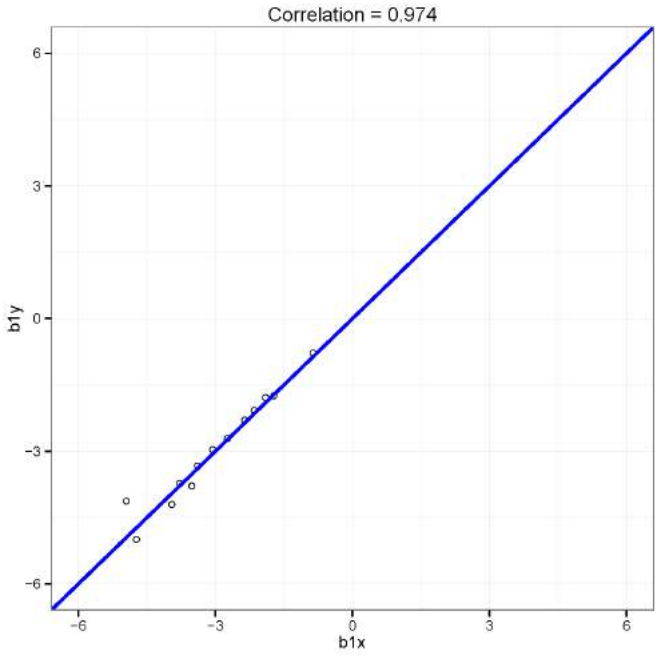


Figure C.44: Grade 10 Math Difficulty Parameter (b1) for Items with Four Score Categories

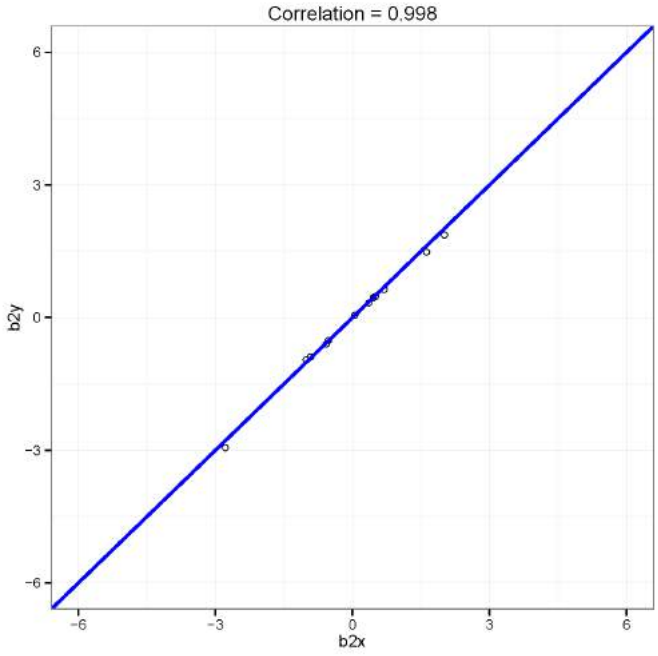


Figure C.45: Grade 10 Math Difficulty Parameter (b2) for Items with Four Score Categories

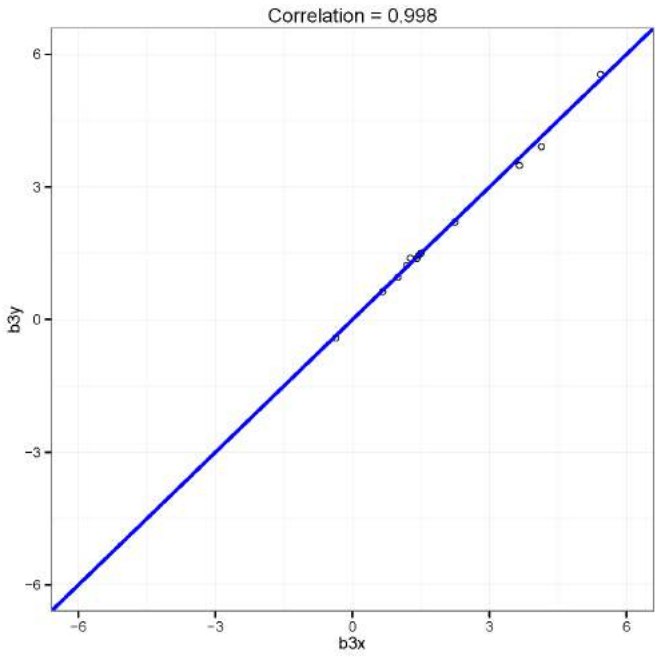


Figure C.46: Grade 10 Math Difficulty Parameter (b3) for Items with Four Score Categories

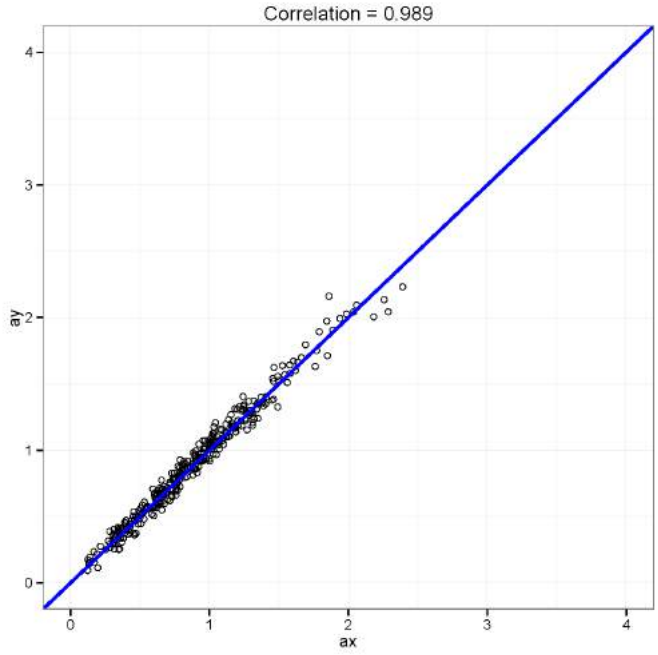


Figure C.47: Grade 3 ELA Discrimination Parameter for All Items

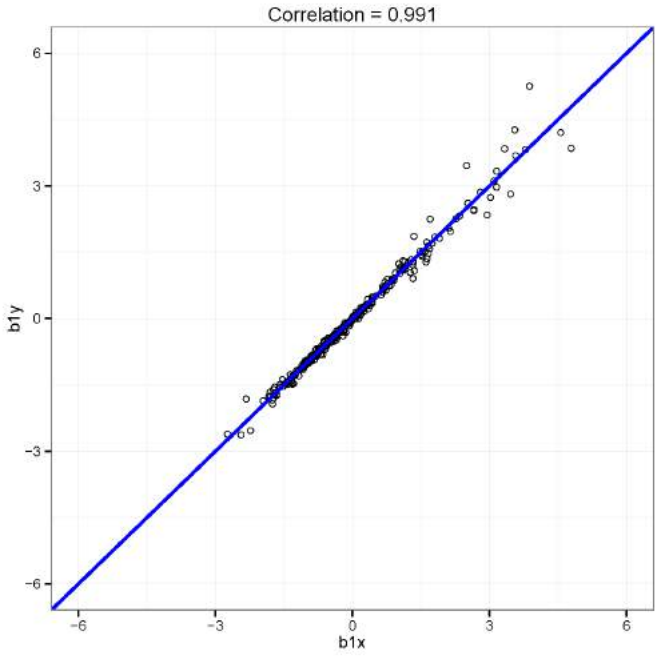


Figure C.48: Grade 3 ELA Difficulty Parameter (b1) for Items with Two Score Categories

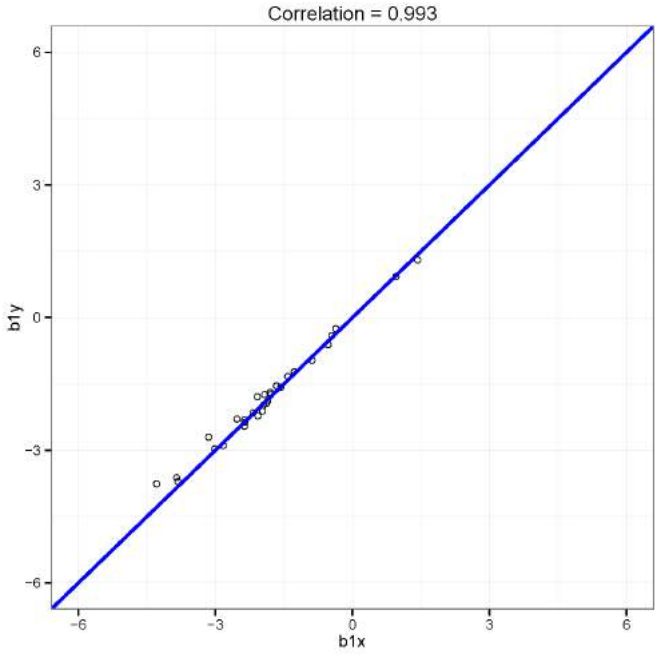


Figure C.49: Grade 3 ELA Difficulty Parameter (b1) for Items with Three Score Categories

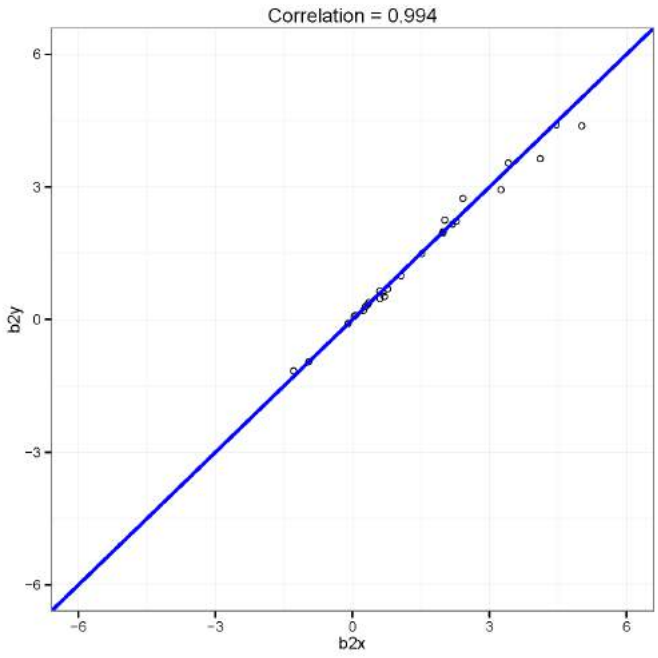


Figure C.50: Grade 3 ELA Difficulty Parameter (b2) for Items with Three Score Categories

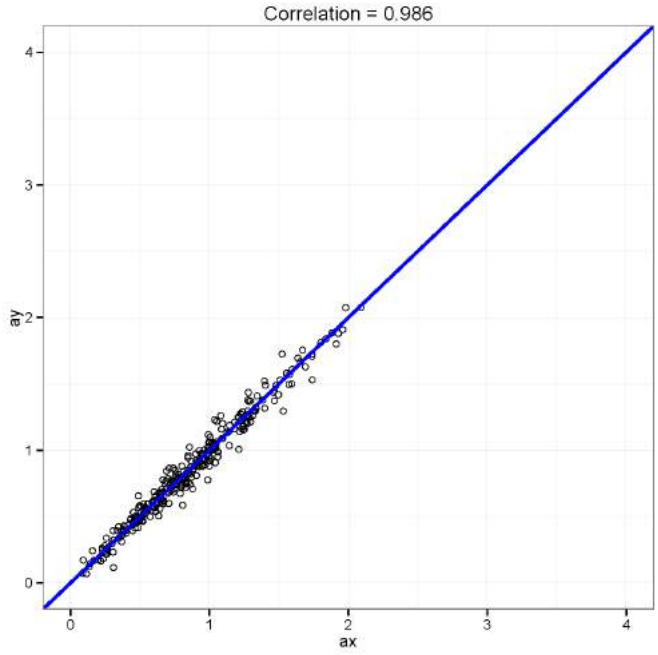


Figure C.51: Grade 4 ELA Discrimination Parameter for All Items

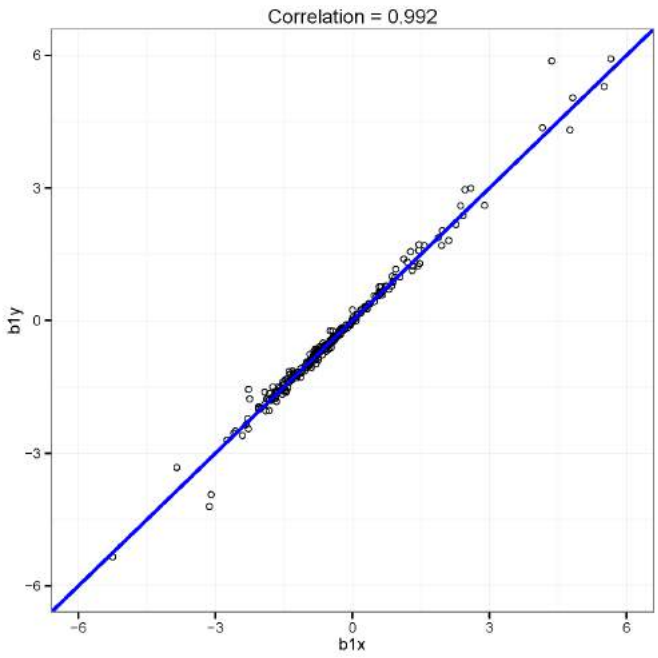


Figure C.52: Grade 4 ELA Difficulty Parameter (b1) for Items with Two Score Categories

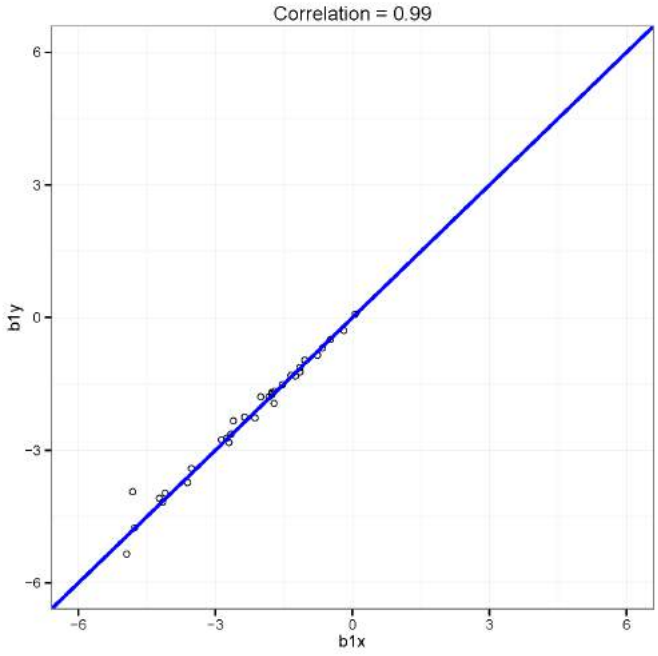


Figure C.53: Grade 4 ELA Difficulty Parameter (b1) for Items with Three Score Categories

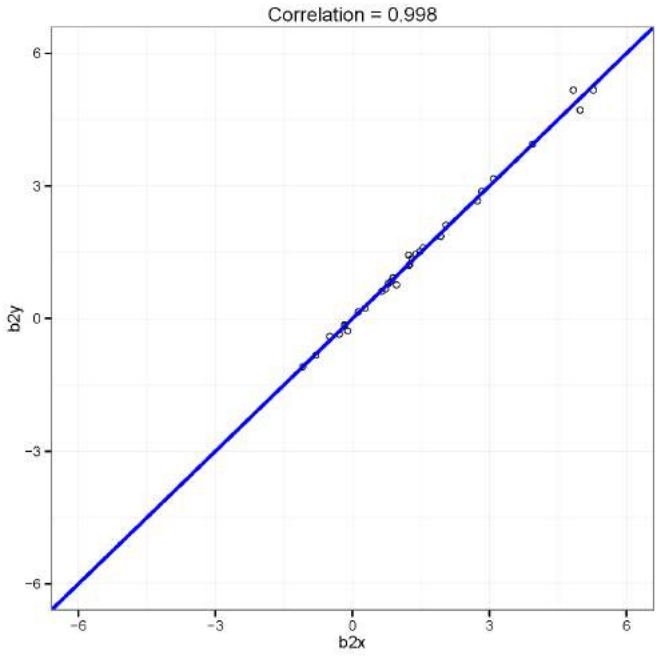


Figure C.54: Grade 4 ELA Difficulty Parameter (b2) for Items with Three Score Categories

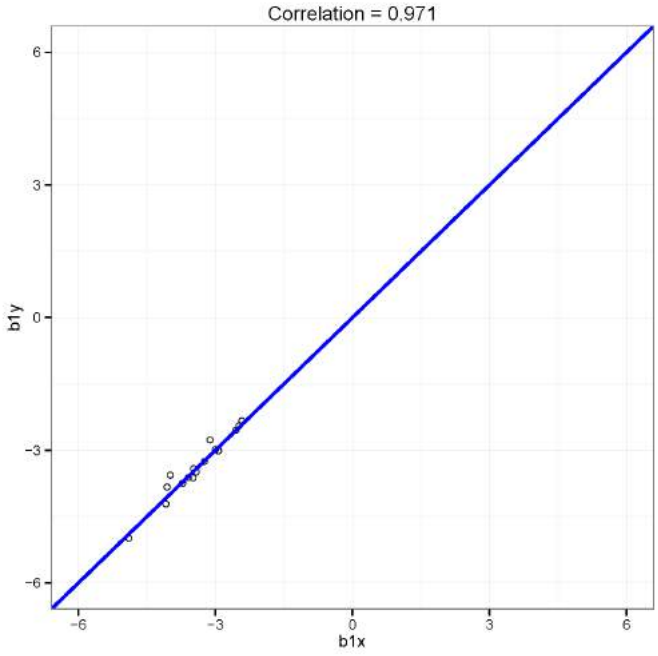


Figure C.55: Grade 4 ELA Difficulty Parameter (b1) for Items with Four Score Categories

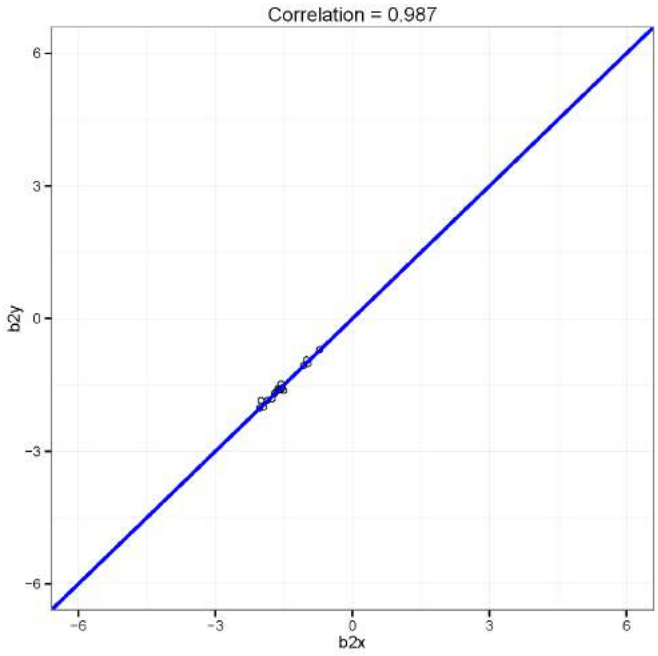


Figure C.56: Grade 4 ELA Difficulty Parameter (b2) for Items with Four Score Categories

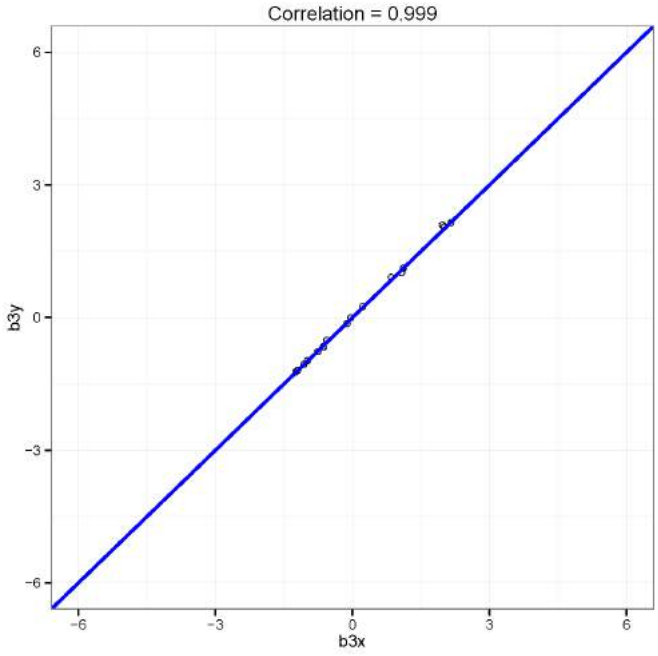


Figure C.57: Grade 4 ELA Difficulty Parameter (b3) for Items with Four Score Categories

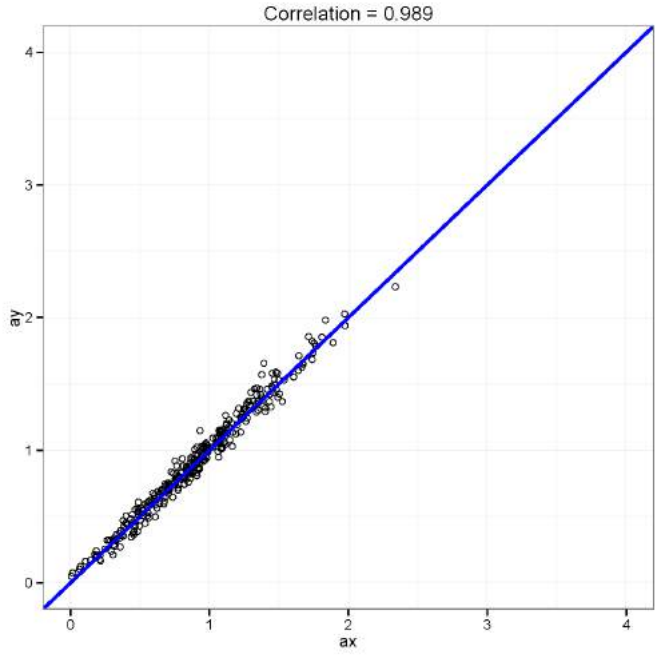


Figure C.58: Grade 5 ELA Discrimination Parameter for All Items

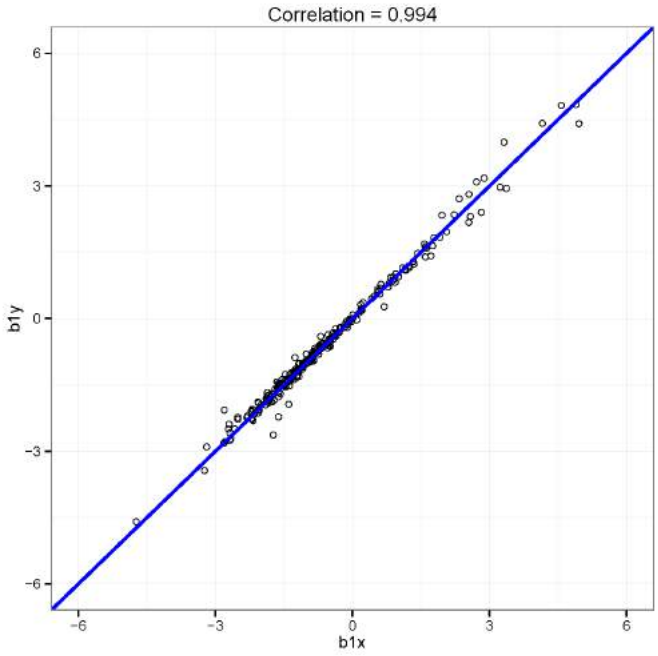


Figure C.59: Grade 5 ELA Difficulty Parameter (b1) for Items with Two Score Categories

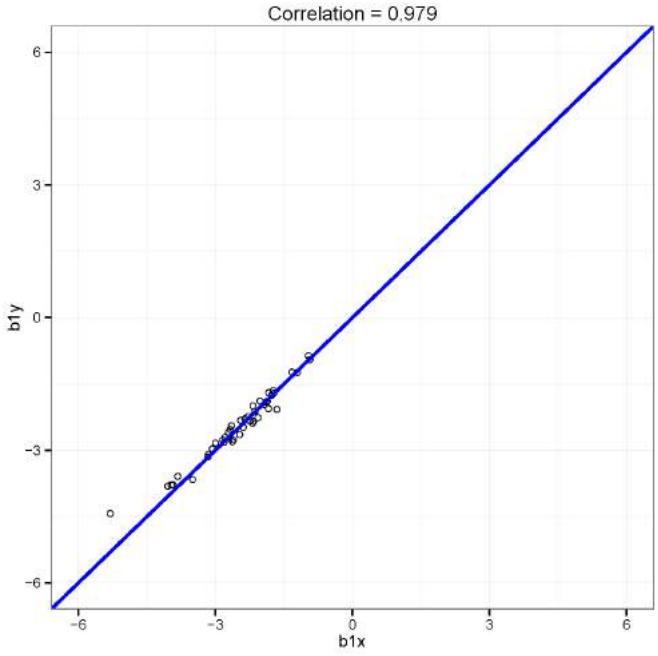


Figure C.60: Grade 5 ELA Difficulty Parameter (b1) for Items with Three Score Categories

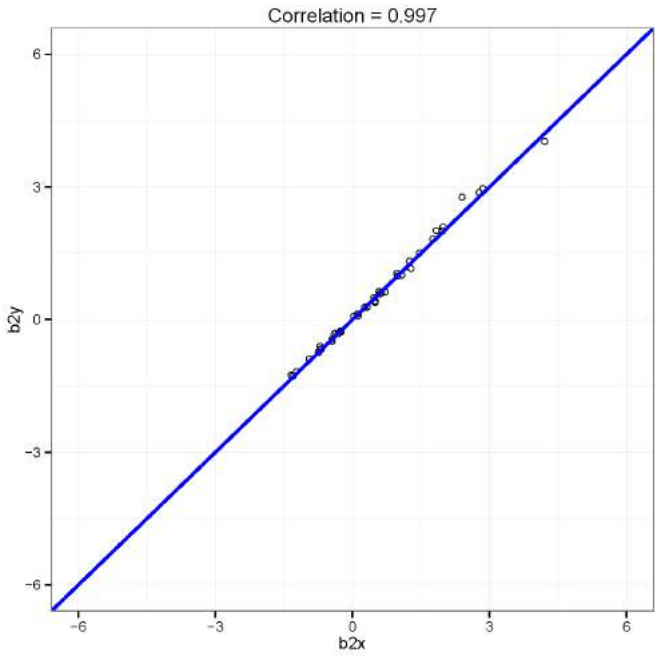


Figure C.61: Grade 5 ELA Difficulty Parameter (b2) for Items with Three Score Categories

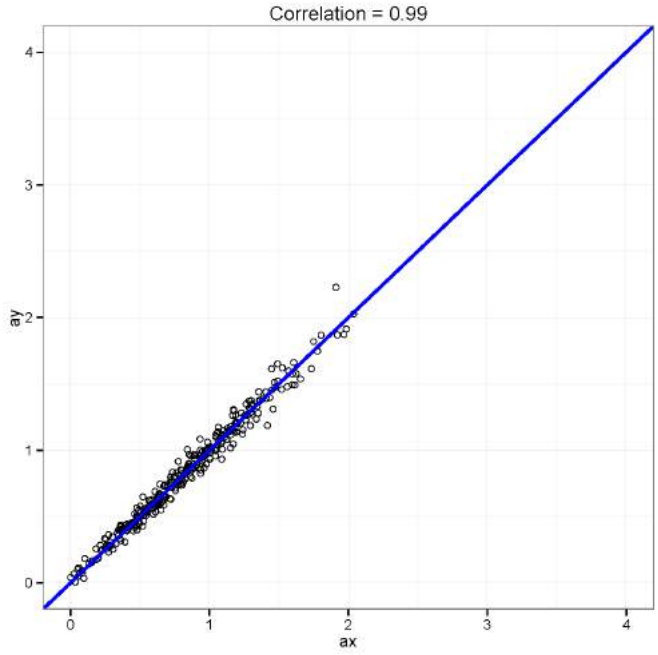


Figure C.62: Grade 6 ELA Discrimination Parameter for All Items

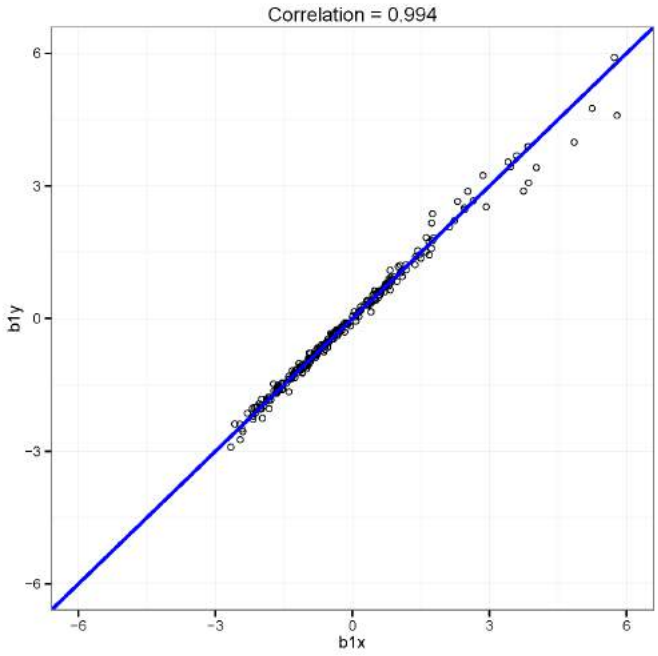


Figure C.63: Grade 6 ELA Difficulty Parameter (b1) for Items with Two Score Categories

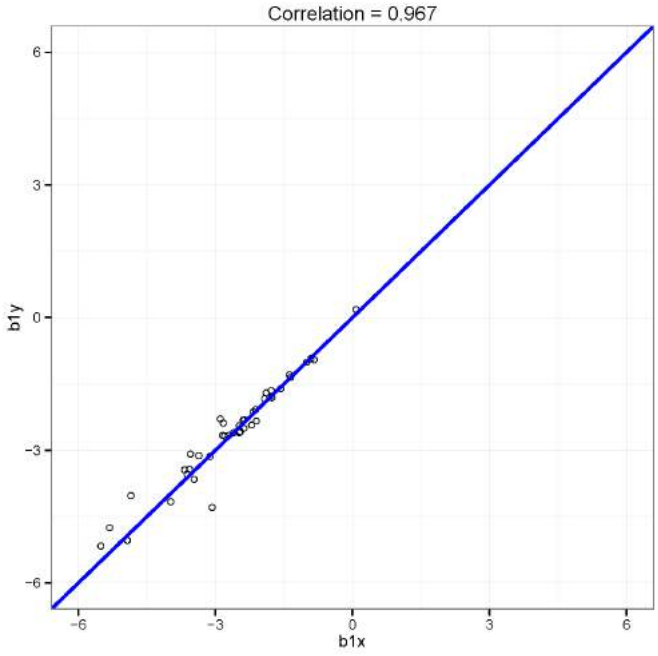


Figure C.64: Grade 6 ELA Difficulty Parameter (b1) for Items with Three Score Categories

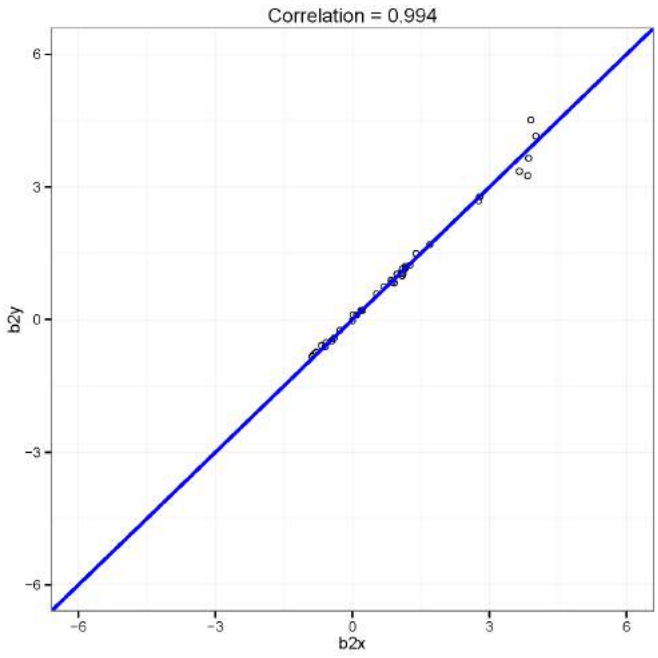


Figure C.65: Grade 6 ELA Difficulty Parameter (b2) for Items with Three Score Categories

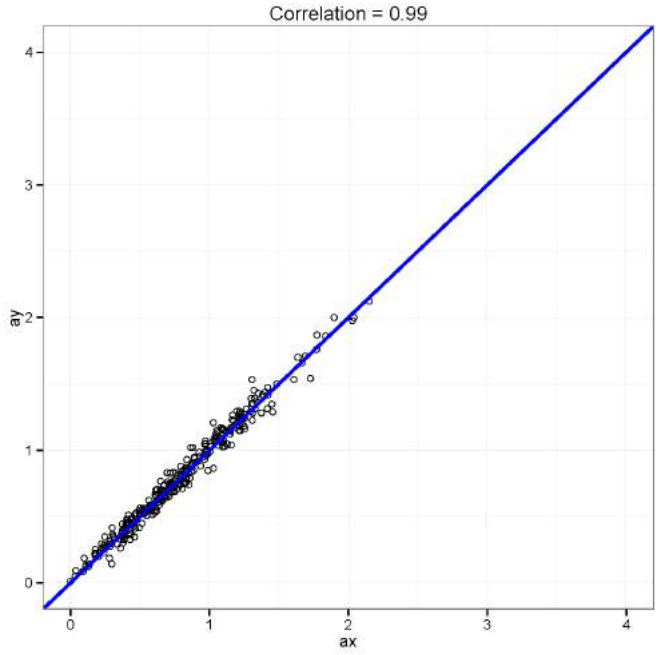


Figure C.66: Grade 7 ELA Discrimination Parameter for All Items

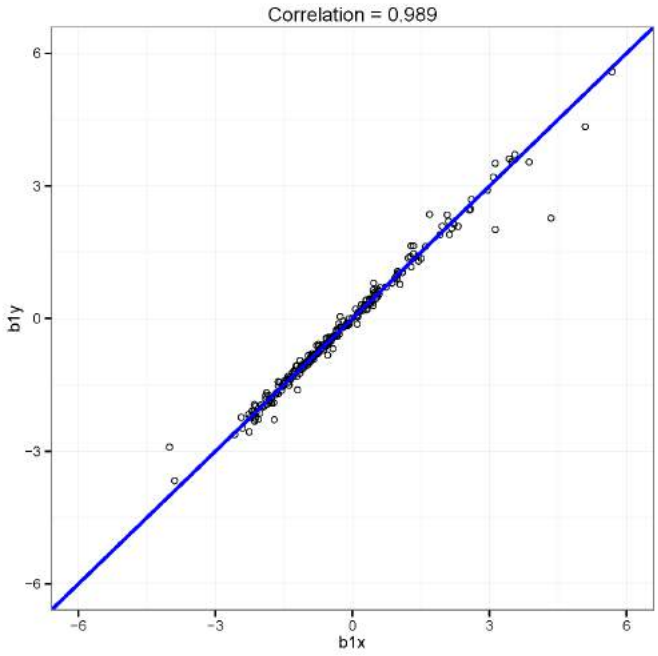


Figure C.67: Grade 7 ELA Difficulty Parameter (b1) for Items with Two Score Categories

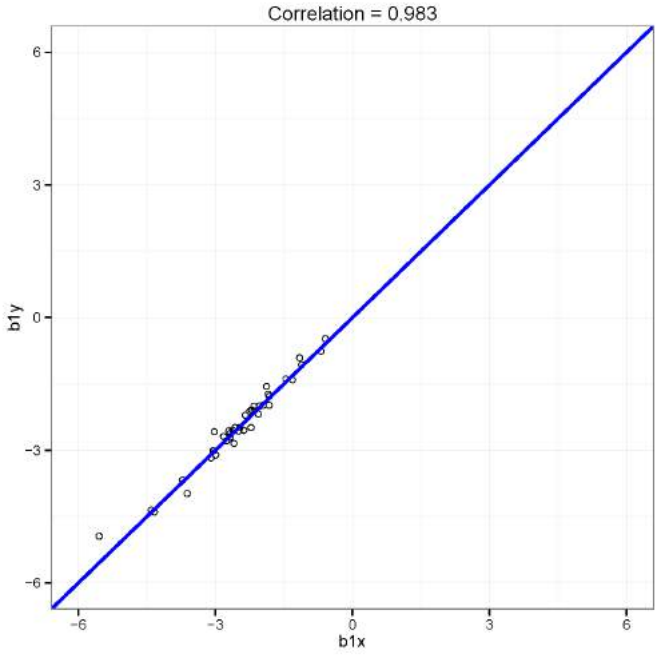


Figure C.68: Grade 7 ELA Difficulty Parameter (b1) for Items with Three Score Categories

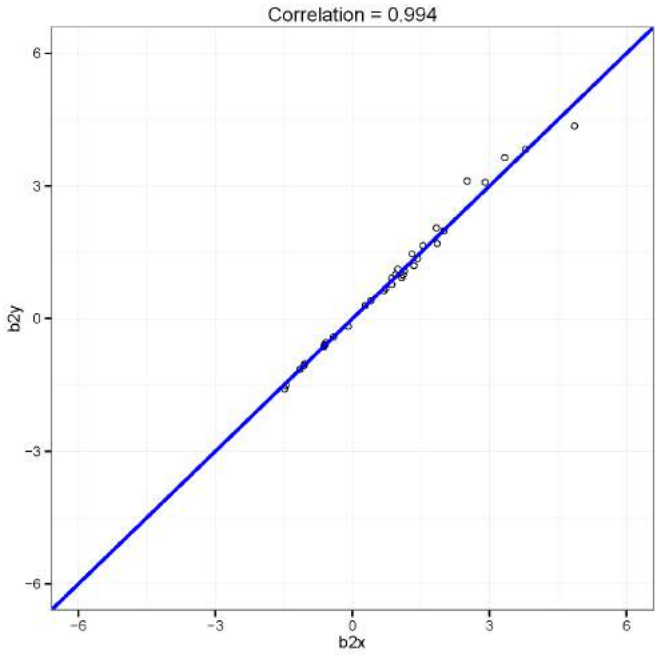


Figure C.69: Grade 7 ELA Difficulty Parameter (b2) for Items with Three Score Categories

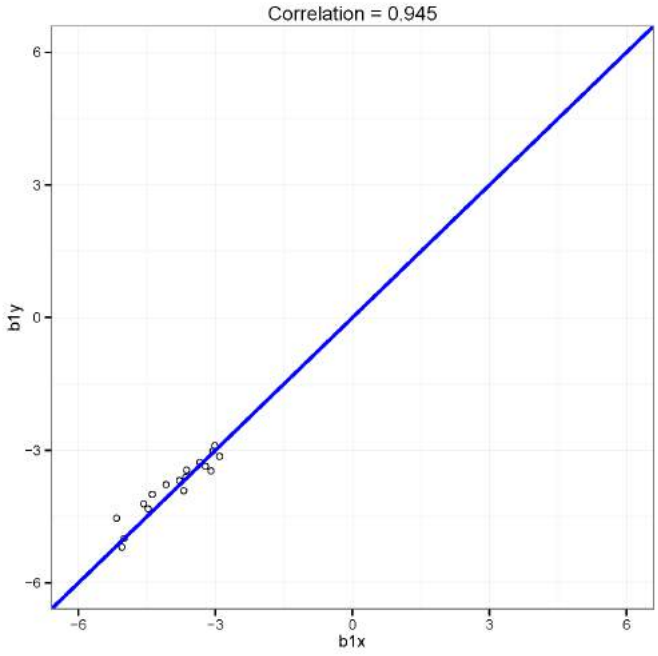


Figure C.70: Grade 7 ELA Difficulty Parameter (b1) for Items with Four Score Categories

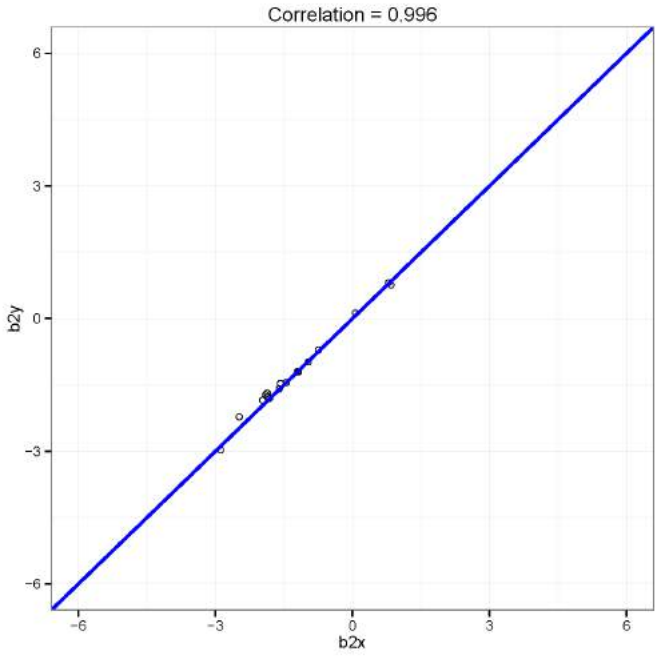


Figure C.71: Grade 7 ELA Difficulty Parameter (b2) for Items with Four Score Categories

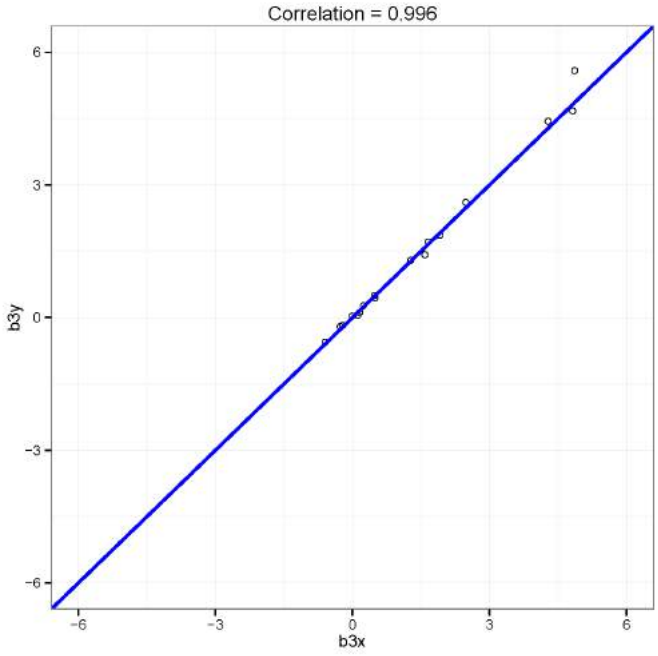


Figure C.72: Grade 7 ELA Difficulty Parameter (b3) for Items with Four Score Categories

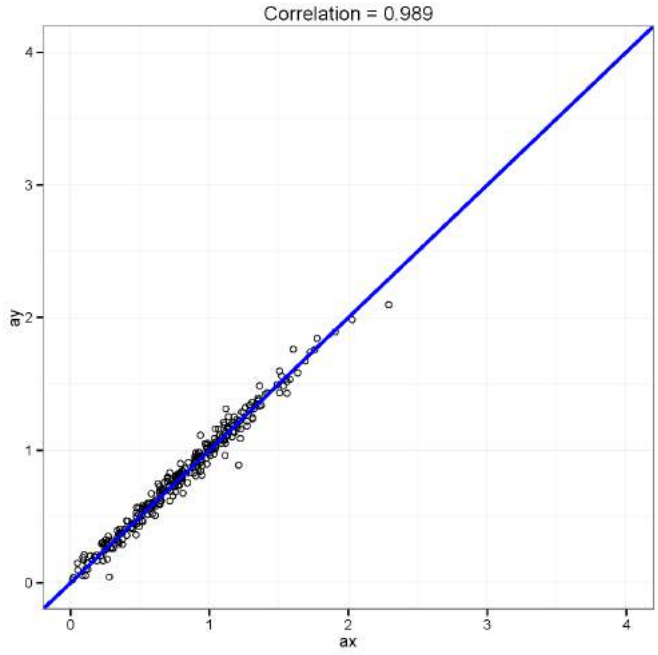


Figure C.73: Grade 8 ELA Discrimination Parameter for All Items

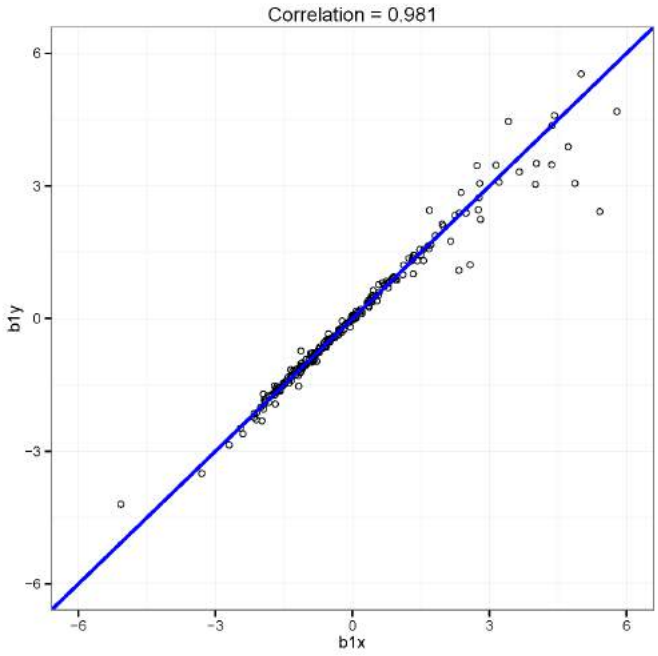


Figure C.74: Grade 8 ELA Difficulty Parameter (b1) for Items with Two Score Categories

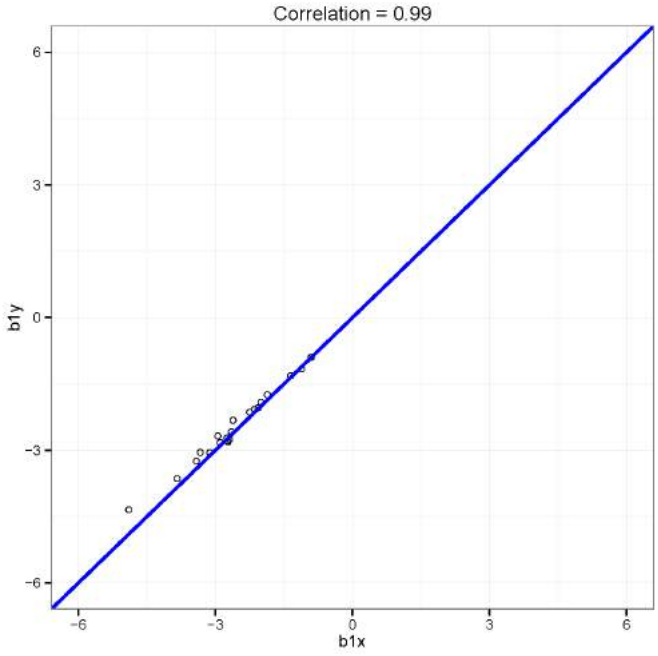


Figure C.75: Grade 8 ELA Difficulty Parameter (b1) for Items with Three Score Categories

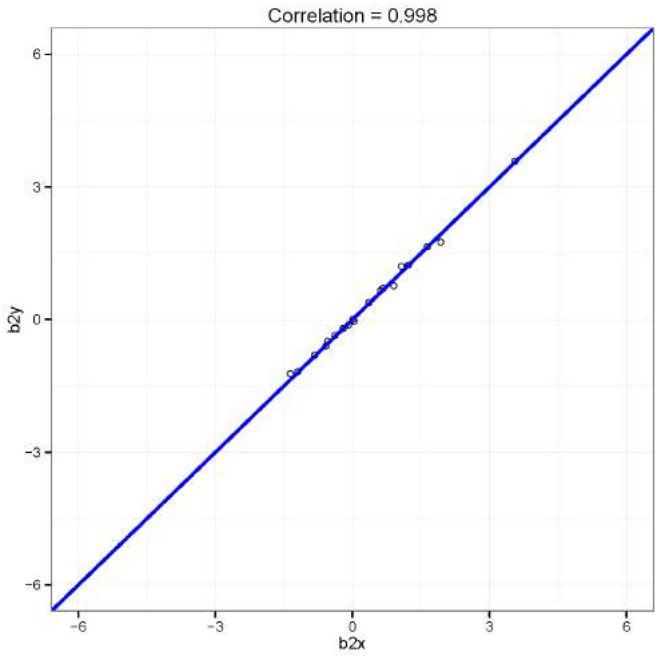


Figure C.76: Grade 8 ELA Difficulty Parameter (b2) for Items with Three Score Categories

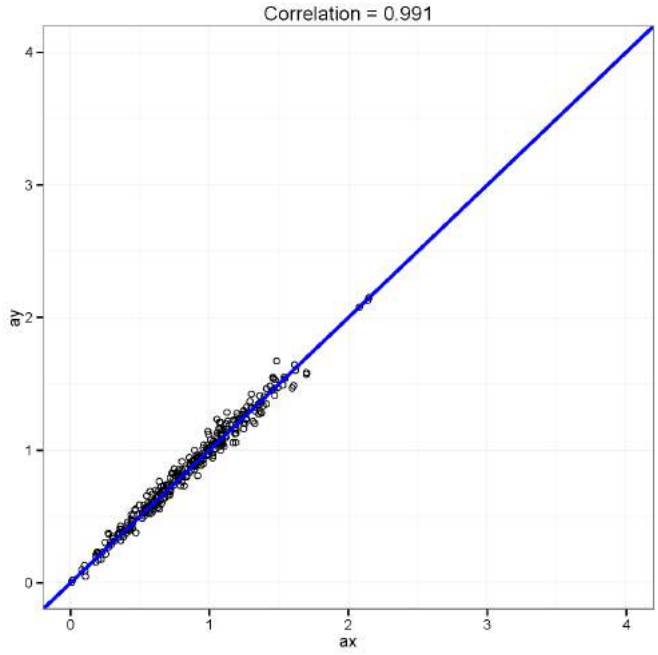


Figure C.77: Grade 10 ELA Discrimination Parameter for All Items

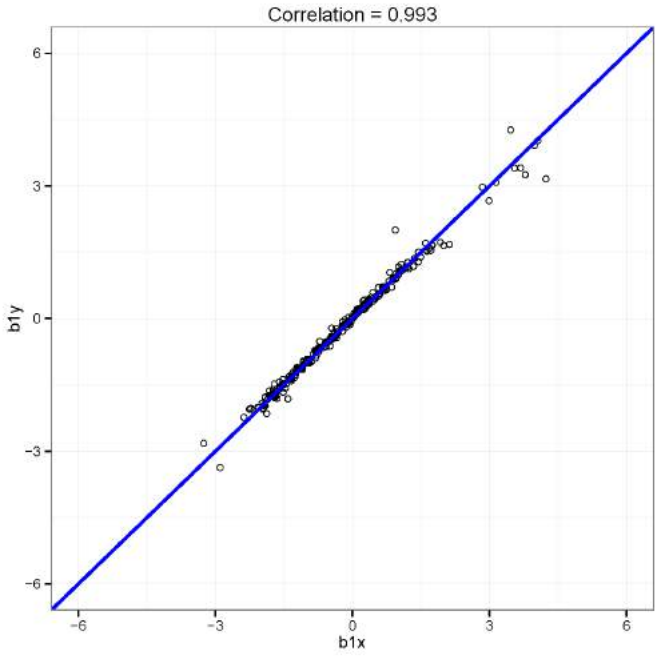


Figure C.78: Grade 10 ELA Difficulty Parameter (b1) for Items with Two Score Categories

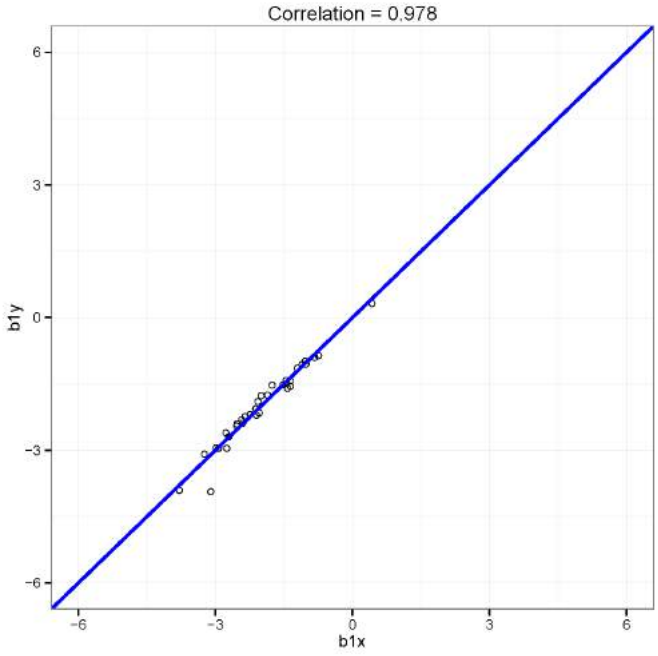


Figure C.79: Grade 10 ELA Difficulty Parameter (b1) for Items with Three Score Categories

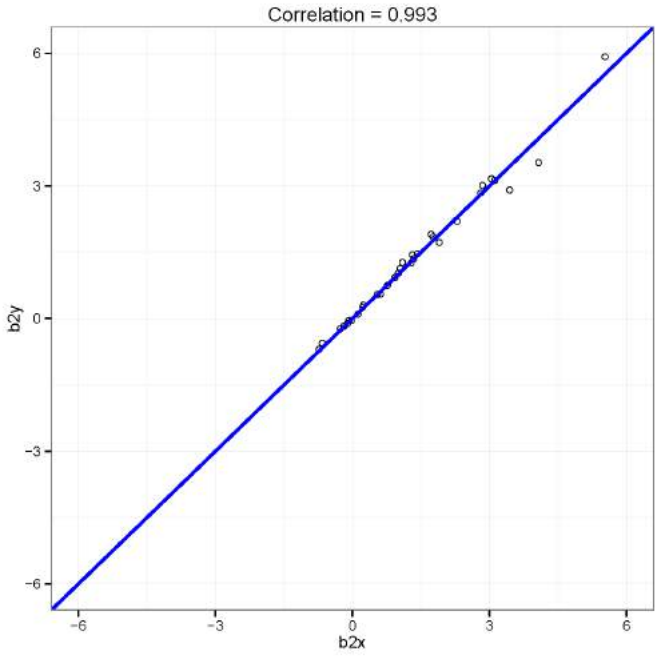


Figure C.80: Grade 10 ELA Difficulty Parameter (b2) for Items with Three Score Categories

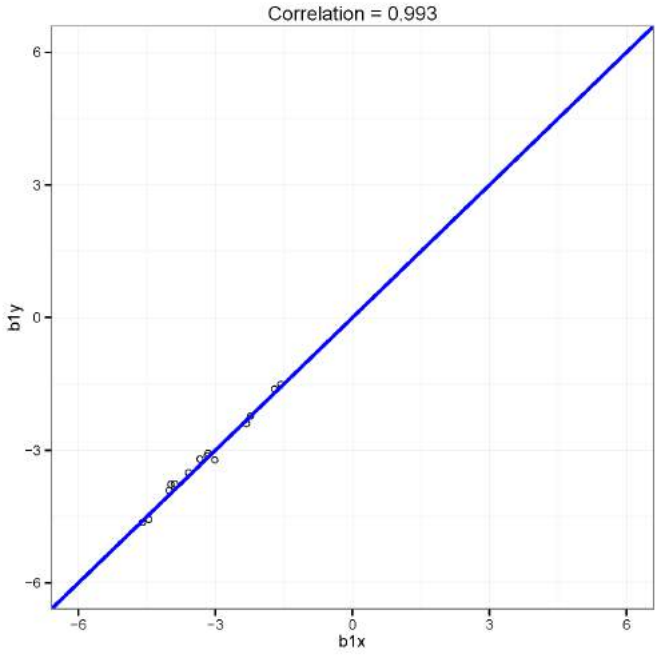


Figure C.81: Grade 10 ELA Difficulty Parameter (b1) for Items with Four Score Categories

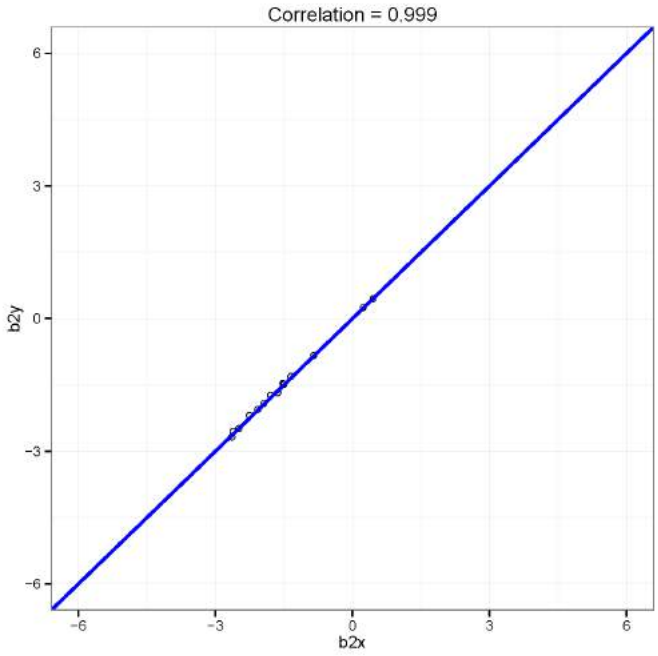


Figure C.82: Grade 10 ELA Difficulty Parameter (b2) for Items with Four Score Categories

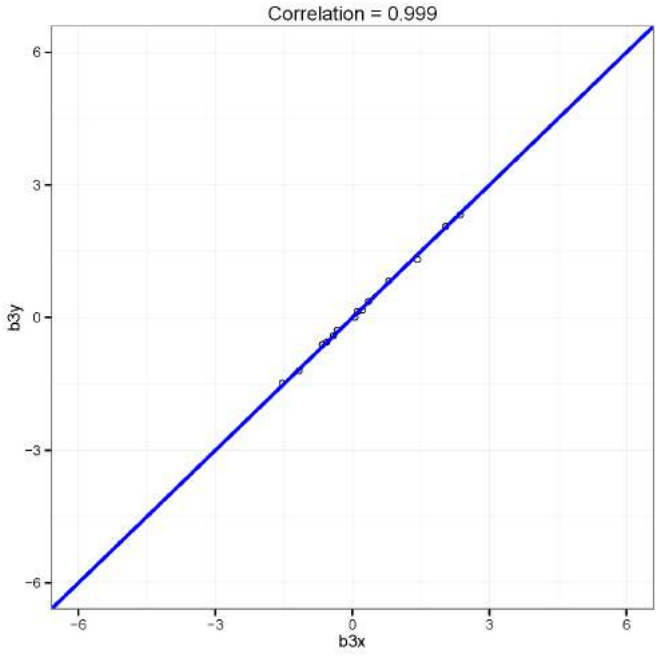


Figure C.83: Grade 10 ELA Difficulty Parameter (b3) for Items with Four Score Categories

D
Subgroup Reliability

Groups with less than 100 students are surpressed in the following tables.

Table D.1: Marginal Scaled Score Reliability for Grade 3 Math Subgroups for Form A

Form	Group	N Population	N	Max Score	Reliability
A	AfricanAmerican	7982	677	64	0.91
A	AmericanIndian	7982	299	64	0.91
A	Asian	7982	199	64	0.91
A	White	7982	6304	64	0.91
A	Hispanic	7982	1833	64	0.91
A	SWD	7982	2466	64	0.91
A	ESOL	7982	1350	64	0.91

Table D.2: Marginal Scaled Score Reliability for Grade 3 Math Subgroups for Form B

Form	Group	N Population	N	Max Score	Reliability
B	AfricanAmerican	4176	307	51	0.90
B	AmericanIndian	4176	138	51	0.90
B	Asian	4176	125	51	0.88
B	White	4176	3355	51	0.90
B	Hispanic	4176	819	51	0.90
B	SWD	4176	300	51	0.90
B	ESOL	4176	510	51	0.90

Table D.3: Marginal Scaled Score Reliability for Grade 3 Math Subgroups for Form C

Form	Group	N Population	N	Max Score	Reliability
C	AfricanAmerican	4178	285	63	0.92
C	AmericanIndian	4178	122	63	0.92
C	Asian	4178	109	63	0.90
C	White	4178	3395	63	0.91
C	Hispanic	4178	794	63	0.92
C	SWD	4178	305	63	0.92
C	ESOL	4178	492	63	0.92

Table D.4: Marginal Scaled Score Reliability for Grade 3 Math Subgroups for Form D

Form	Group	N Population	N	Max Score	Reliability
D	AfricanAmerican	4166	317	66	0.93
D	AmericanIndian	4166	106	66	0.93
D	Asian	4166	119	66	0.91
D	White	4166	3339	66	0.92
D	Hispanic	4166	759	66	0.93
D	SWD	4166	296	66	0.92
D	ESOL	4166	486	66	0.93

Table D.5: Marginal Scaled Score Reliability for Grade 3 Math Subgroups for Form E

Form	Group	N Population	N	Max Score	Reliability
E	AfricanAmerican	4017	277	52	0.90
E	AmericanIndian	4017	112	52	0.90
E	Asian	4017	116	52	0.90
E	White	4017	3256	52	0.90
E	Hispanic	4017	748	52	0.91
E	SWD	4017	250	52	0.90
E	ESOL	4017	479	52	0.91

Table D.6: Marginal Scaled Score Reliability for Grade 3 Math Subgroups for Form F

Form	Group	N Population	N	Max Score	Reliability
F	AfricanAmerican	4182	316	61	0.92
F	AmericanIndian	4182	119	61	0.92
F	Asian	4182	127	61	0.91
F	White	4182	3343	61	0.92
F	Hispanic	4182	800	61	0.92
F	SWD	4182	322	61	0.92
F	ESOL	4182	531	61	0.92

Table D.7: Marginal Scaled Score Reliability for Grade 3 Math Subgroups for Form G

Form	Group	N Population	N	Max Score	Reliability
G	AfricanAmerican	4202	324	64	0.92
G	AmericanIndian	4202	106	64	0.92
G	Asian	4202	142	64	0.91
G	White	4202	3351	64	0.92
G	Hispanic	4202	742	64	0.92
G	SWD	4202	307	64	0.92
G	ESOL	4202	464	64	0.92

Table D.8: Marginal Scaled Score Reliability for Grade 3 Math Subgroups for Form H

Form	Group	N Population	N	Max Score	Reliability
H	AfricanAmerican	4078	284	63	0.93
H	AmericanIndian	4078	125	63	0.93
H	Asian	4078	110	63	0.91
H	White	4078	3288	63	0.92
H	Hispanic	4078	753	63	0.93
H	SWD	4078	289	63	0.92
H	ESOL	4078	471	63	0.93

Table D.9: Marginal Scaled Score Reliability for Grade 4 Math Subgroups for Form A

Form	Group	N Population	N	Max Score	Reliability
A	AfricanAmerican	8272	700	67	0.93
A	AmericanIndian	8272	394	67	0.93
A	Asian	8272	189	67	0.91
A	White	8272	6426	67	0.92
A	Hispanic	8272	1914	67	0.93
A	SWD	8272	2738	67	0.93
A	ESOL	8272	1431	67	0.93

Table D.10: Marginal Scaled Score Reliability for Grade 4 Math Subgroups by Form B

Form	Group	N Population	N	Max Score	Reliability
B	AfricanAmerican	3999	290	46	0.91
B	AmericanIndian	3999	122	46	0.91
B	Asian	3999	121	46	0.88
B	White	3999	3209	46	0.90
B	Hispanic	3999	719	46	0.91
B	SWD	3999	269	46	0.90
B	ESOL	3999	474	46	0.91

Table D.11: Marginal Scaled Score Reliability for Grade 4 Math Subgroups by Form C

Form	Group	N Population	N	Max Score	Reliability
C	AfricanAmerican	4017	237	67	0.92
C	AmericanIndian	4017	131	67	0.92
C	Asian	4017	135	67	0.90
C	White	4017	3244	67	0.92
C	Hispanic	4017	759	67	0.92
C	SWD	4017	260	67	0.92
C	ESOL	4017	494	67	0.92

Table D.12: Marginal Scaled Score Reliability for Grade 4 Math Subgroups by Form D

Form	Group	N Population	N	Max Score	Reliability
D	AfricanAmerican	4146	291	66	0.93
D	AmericanIndian	4146	122	66	0.93
D	Asian	4146	122	66	0.92
D	White	4146	3349	66	0.93
D	Hispanic	4146	724	66	0.93
D	SWD	4146	261	66	0.93
D	ESOL	4146	456	66	0.93

Table D.13: Marginal Scaled Score Reliability for Grade 4 Math Subgroups by Form E

Form	Group	N Population	N	Max Score	Reliability
E	AfricanAmerican	4020	273	51	0.91
E	AmericanIndian	4020	150	51	0.92
E	Asian	4020	121	51	0.90
E	White	4020	3197	51	0.91
E	Hispanic	4020	774	51	0.91
E	SWD	4020	231	51	0.91
E	ESOL	4020	481	51	0.92

Table D.14: Marginal Scaled Score Reliability for Grade 4 Math Subgroups by Form F

Form	Group	N Population	N	Max Score	Reliability
F	AfricanAmerican	4065	302	68	0.93
F	AmericanIndian	4065	139	68	0.93
F	Asian	4065	106	68	0.92
F	White	4065	3274	68	0.93
F	Hispanic	4065	721	68	0.93
F	SWD	4065	260	68	0.93
F	ESOL	4065	447	68	0.93

Table D.15: Marginal Scaled Score Reliability for Grade 4 Math Subgroups by Form G

Form	Group	N Population	N	Max Score	Reliability
G	AfricanAmerican	4009	312	67	0.93
G	AmericanIndian	4009	143	67	0.93
G	Asian	4009	126	67	0.91
G	White	4009	3195	67	0.93
G	Hispanic	4009	749	67	0.93
G	SWD	4009	264	67	0.93
G	ESOL	4009	472	67	0.93

Table D.16: Marginal Scaled Score Reliability for Grade 4 Math Subgroups by Form H

Form	Group	N Population	N	Max Score	Reliability
H	AfricanAmerican	3946	245	64	0.93
H	AmericanIndian	3946	142	64	0.93
H	Asian	3946	112	64	0.91
H	White	3946	3184	64	0.92
H	Hispanic	3946	724	64	0.93
H	SWD	3946	260	64	0.93
H	ESOL	3946	466	64	0.93

Table D.17: Marginal Scaled Score Reliability for Grade 5 Math Subgroups by Form A

Form	Group	N Population	N	Max Score	Reliability
A	AfricanAmerican	8027	674	68	0.93
A	AmericanIndian	8027	408	68	0.93
A	Asian	8027	224	68	0.92
A	White	8027	6187	68	0.93
A	Hispanic	8027	1776	68	0.93
A	SWD	8027	2748	68	0.92
A	ESOL	8027	1346	68	0.93

Table D.18: Marginal Scaled Score Reliability for Grade 5 Math Subgroups by Form B

Form	Group	N Population	N	Max Score	Reliability
B	AfricanAmerican	4018	281	52	0.91
B	AmericanIndian	4018	172	52	0.91
B	Asian	4018	103	52	0.89
B	White	4018	3214	52	0.90
B	Hispanic	4018	745	52	0.91
B	SWD	4018	246	52	0.90
B	ESOL	4018	490	52	0.91

Table D.19: Marginal Scaled Score Reliability for Grade 5 Math Subgroups by Form C

Form	Group	N Population	N	Max Score	Reliability
C	AfricanAmerican	4024	272	65	0.93
C	AmericanIndian	4024	149	65	0.93
C	Asian	4024	122	65	0.91
C	White	4024	3218	65	0.92
C	Hispanic	4024	756	65	0.93
C	SWD	4024	252	65	0.92
C	ESOL	4024	474	65	0.93

Table D.20: Marginal Scaled Score Reliability for Grade 5 Math Subgroups by Form D

Form	Group	N Population	N	Max Score	Reliability
D	AfricanAmerican	4031	252	65	0.92
D	AmericanIndian	4031	162	65	0.92
D	Asian	4031	118	65	0.91
D	White	4031	3221	65	0.92
D	Hispanic	4031	753	65	0.92
D	SWD	4031	227	65	0.92
D	ESOL	4031	443	65	0.92

Table D.21: Marginal Scaled Score Reliability for Grade 5 Math Subgroups by Form E

Form	Group	N Population	N	Max Score	Reliability
E	AfricanAmerican	3996	291	51	0.90
E	AmericanIndian	3996	141	51	0.91
E	Asian	3996	123	51	0.89
E	White	3996	3172	51	0.91
E	Hispanic	3996	708	51	0.91
E	SWD	3996	239	51	0.90
E	ESOL	3996	471	51	0.91

Table D.22: Marginal Scaled Score Reliability for Grade 5 Math Subgroups by Form F

Form	Group	N Population	N	Max Score	Reliability
F	AfricanAmerican	4037	283	62	0.92
F	AmericanIndian	4037	151	62	0.92
F	Asian	4037	136	62	0.90
F	White	4037	3214	62	0.92
F	Hispanic	4037	743	62	0.92
F	SWD	4037	241	62	0.92
F	ESOL	4037	495	62	0.92

Table D.23: Marginal Scaled Score Reliability for Grade 5 Math Subgroups by Form G

Form	Group	N Population	N	Max Score	Reliability
G	AfricanAmerican	4006	254	64	0.93
G	AmericanIndian	4006	139	64	0.93
G	Asian	4006	132	64	0.90
G	White	4006	3198	64	0.92
G	Hispanic	4006	730	64	0.93
G	SWD	4006	235	64	0.92
G	ESOL	4006	468	64	0.93

Table D.24: Marginal Scaled Score Reliability for Grade 5 Math Subgroups by Form H

Form	Group	N Population	N	Max Score	Reliability
H	AfricanAmerican	3968	241	65	0.92
H	AmericanIndian	3968	139	65	0.93
H	Asian	3968	110	65	0.92
H	White	3968	3230	65	0.92
H	Hispanic	3968	739	65	0.92
H	SWD	3968	218	65	0.92
H	ESOL	3968	445	65	0.92

Table D.25: Marginal Scaled Score Reliability for Grade 6 Math Subgroups by Form A

Form	Group	N Population	N	Max Score	Reliability
A	AfricanAmerican	7652	632	69	0.91
A	AmericanIndian	7652	381	69	0.92
A	Asian	7652	192	69	0.88
A	White	7652	5953	69	0.91
A	Hispanic	7652	1634	69	0.91
A	SWD	7652	2609	69	0.91
A	ESOL	7652	1179	69	0.91

Table D.26: Marginal Scaled Score Reliability for Grade 6 Math Subgroups by Form B

Form	Group	N Population	N	Max Score	Reliability
B	AfricanAmerican	4168	312	50	0.89
B	AmericanIndian	4168	168	50	0.89
B	Asian	4168	123	50	0.85
B	White	4168	3307	50	0.88
B	Hispanic	4168	745	50	0.89
B	SWD	4168	248	50	0.88
B	ESOL	4168	483	50	0.89

Table D.27: Marginal Scaled Score Reliability for Grade 6 Math Subgroups by Form C

Form	Group	N Population	N	Max Score	Reliability
C	AfricanAmerican	4154	278	64	0.91
C	AmericanIndian	4154	156	64	0.91
C	Asian	4154	131	64	0.88
C	White	4154	3352	64	0.90
C	Hispanic	4154	777	64	0.91
C	SWD	4154	224	64	0.91
C	ESOL	4154	503	64	0.91

Table D.28: Marginal Scaled Score Reliability for Grade 6 Math Subgroups by Form D

Form	Group	N Population	N	Max Score	Reliability
D	AfricanAmerican	4081	286	64	0.92
D	AmericanIndian	4081	179	64	0.92
D	Asian	4081	119	64	0.90
D	White	4081	3262	64	0.91
D	Hispanic	4081	732	64	0.91
D	SWD	4081	265	64	0.91
D	ESOL	4081	468	64	0.92

Table D.29: Marginal Scaled Score Reliability for Grade 6 Math Subgroups by Form E

Form	Group	N Population	N	Max Score	Reliability
E	AfricanAmerican	4124	284	49	0.90
E	AmericanIndian	4124	171	49	0.90
E	Asian	4124	135	49	0.88
E	White	4124	3315	49	0.89
E	Hispanic	4124	697	49	0.90
E	SWD	4124	230	49	0.89
E	ESOL	4124	455	49	0.90

Table D.30: Marginal Scaled Score Reliability for Grade 6 Math Subgroups by Form F

Form	Group	N Population	N	Max Score	Reliability
F	AfricanAmerican	4021	302	58	0.90
F	AmericanIndian	4021	183	58	0.90
F	Asian	4021	119	58	0.86
F	White	4021	3188	58	0.89
F	Hispanic	4021	737	58	0.90
F	SWD	4021	223	58	0.89
F	ESOL	4021	466	58	0.90

Table D.31: Marginal Scaled Score Reliability for Grade 6 Math Subgroups by Form G

Form	Group	N Population	N	Max Score	Reliability
G	AfricanAmerican	4037	284	65	0.92
G	AmericanIndian	4037	160	65	0.92
G	Asian	4037	108	65	0.90
G	White	4037	3228	65	0.92
G	Hispanic	4037	729	65	0.92
G	SWD	4037	212	65	0.92
G	ESOL	4037	427	65	0.92

Table D.32: Marginal Scaled Score Reliability for Grade 6 Math Subgroups by Form H

Form	Group	N Population	N	Max Score	Reliability
H	AfricanAmerican	4172	299	53	0.90
H	AmericanIndian	4172	167	53	0.90
H	Asian	4172	136	53	0.87
H	White	4172	3338	53	0.89
H	Hispanic	4172	766	53	0.90
H	SWD	4172	245	53	0.90
H	ESOL	4172	503	53	0.90

Table D.33: Marginal Scaled Score Reliability for Grade 7 Math Subgroups by Form A

Form	Group	N Population	N	Max Score	Reliability
A	AfricanAmerican	7320	601	60	0.90
A	AmericanIndian	7320	397	60	0.90
A	Asian	7320	175	60	0.89
A	White	7320	5729	60	0.90
A	Hispanic	7320	1589	60	0.90
A	SWD	7320	2563	60	0.90
A	ESOL	7320	1101	60	0.90

Table D.34: Marginal Scaled Score Reliability for Grade 7 Math Subgroups by Form B

Form	Group	N Population	N	Max Score	Reliability
B	AfricanAmerican	4156	320	47	0.88
B	AmericanIndian	4156	179	47	0.89
B	Asian	4156	130	47	0.87
B	White	4156	3291	47	0.88
B	Hispanic	4156	697	47	0.88
B	SWD	4156	241	47	0.88
B	ESOL	4156	440	47	0.88

Table D.35: Marginal Scaled Score Reliability for Grade 7 Math Subgroups by Form C

Form	Group	N Population	N	Max Score	Reliability
C	AfricanAmerican	4099	301	63	0.89
C	AmericanIndian	4099	167	63	0.89
C	Asian	4099	121	63	0.87
C	White	4099	3277	63	0.89
C	Hispanic	4099	708	63	0.89
C	SWD	4099	217	63	0.89
C	ESOL	4099	446	63	0.89

Table D.36: Marginal Scaled Score Reliability for Grade 7 Math Subgroups by Form D

Form	Group	N Population	N	Max Score	Reliability
D	AfricanAmerican	4109	302	62	0.90
D	AmericanIndian	4109	183	62	0.90
D	Asian	4109	116	62	0.89
D	White	4109	3255	62	0.90
D	Hispanic	4109	733	62	0.90
D	SWD	4109	220	62	0.89
D	ESOL	4109	476	62	0.90

Table D.37: Marginal Scaled Score Reliability for Grade 7 Math Subgroups by Form E

Form	Group	N Population	N	Max Score	Reliability
E	AfricanAmerican	4067	318	51	0.88
E	AmericanIndian	4067	159	51	0.89
E	Asian	4067	118	51	0.88
E	White	4067	3217	51	0.88
E	Hispanic	4067	692	51	0.88
E	SWD	4067	215	51	0.87
E	ESOL	4067	448	51	0.88

Table D.38: Marginal Scaled Score Reliability for Grade 7 Math Subgroups by Form F

Form	Group	N Population	N	Max Score	Reliability
F	AfricanAmerican	4064	288	65	0.90
F	AmericanIndian	4064	152	65	0.90
F	Asian	4064	113	65	0.89
F	White	4064	3242	65	0.90
F	Hispanic	4064	690	65	0.90
F	SWD	4064	232	65	0.90
F	ESOL	4064	432	65	0.90

Table D.39: Marginal Scaled Score Reliability for Grade 7 Math Subgroups by Form G

Form	Group	N Population	N	Max Score	Reliability
G	AfricanAmerican	4123	291	61	0.89
G	AmericanIndian	4123	173	61	0.90
G	Asian	4123	115	61	0.89
G	White	4123	3297	61	0.90
G	Hispanic	4123	708	61	0.90
G	SWD	4123	208	61	0.89
G	ESOL	4123	421	61	0.89

Table D.40: Marginal Scaled Score Reliability for Grade 7 Math Subgroups by Form H

Form	Group	N Population	N	Max Score	Reliability
H	AfricanAmerican	4143	270	48	0.89
H	AmericanIndian	4143	161	48	0.89
H	Asian	4143	140	48	0.88
H	White	4143	3327	48	0.89
H	Hispanic	4143	734	48	0.89
H	SWD	4143	255	48	0.88
H	ESOL	4143	435	48	0.89

Table D.41: Marginal Scaled Score Reliability for Grade 8 Math Subgroups by Form A

Form	Group	N Population	N	Max Score	Reliability
A	AfricanAmerican	6849	567	63	0.89
A	AmericanIndian	6849	347	63	0.90
A	Asian	6849	173	63	0.88
A	White	6849	5369	63	0.89
A	Hispanic	6849	1411	63	0.90
A	SWD	6849	2295	63	0.89
A	ESOL	6849	938	63	0.90

Table D.42: Marginal Scaled Score Reliability for Grade 8 Math Subgroups by Form B

Form	Group	N Population	N	Max Score	Reliability
B	AfricanAmerican	4102	286	49	0.87
B	AmericanIndian	4102	169	49	0.87
B	Asian	4102	130	49	0.85
B	White	4102	3274	49	0.87
B	Hispanic	4102	733	49	0.87
B	SWD	4102	215	49	0.86
B	ESOL	4102	425	49	0.87

Table D.43: Marginal Scaled Score Reliability for Grade 8 Math Subgroups by Form C

Form	Group	N Population	N	Max Score	Reliability
C	AfricanAmerican	4151	273	56	0.89
C	AmericanIndian	4151	163	56	0.89
C	Asian	4151	128	56	0.87
C	White	4151	3327	56	0.89
C	Hispanic	4151	717	56	0.89
C	SWD	4151	231	56	0.88
C	ESOL	4151	428	56	0.89

Table D.44: Marginal Scaled Score Reliability for Grade 8 Math Subgroups by Form D

Form	Group	N Population	N	Max Score	Reliability
D	AfricanAmerican	4153	307	61	0.90
D	AmericanIndian	4153	171	61	0.90
D	Asian	4153	134	61	0.88
D	White	4153	3305	61	0.90
D	Hispanic	4153	734	61	0.90
D	SWD	4153	267	61	0.89
D	ESOL	4153	416	61	0.90

Table D.45: Marginal Scaled Score Reliability for Grade 8 Math Subgroups by Form E

Form	Group	N Population	N	Max Score	Reliability
E	AfricanAmerican	4098	286	48	0.87
E	AmericanIndian	4098	179	48	0.87
E	Asian	4098	110	48	0.85
E	White	4098	3270	48	0.87
E	Hispanic	4098	728	48	0.87
E	SWD	4098	214	48	0.86
E	ESOL	4098	427	48	0.87

Table D.46: Marginal Scaled Score Reliability for Grade 8 Math Subgroups by Form G

Form	Group	N Population	N	Max Score	Reliability
G	AfricanAmerican	4083	256	61	0.89
G	AmericanIndian	4083	171	61	0.90
G	Asian	4083	101	61	0.89
G	White	4083	3290	61	0.90
G	Hispanic	4083	712	61	0.89
G	SWD	4083	221	61	0.89
G	ESOL	4083	396	61	0.89

Table D.47: Marginal Scaled Score Reliability for Grade 8 Math Subgroups by Form H

Form	Group	N Population	N	Max Score	Reliability
H	AfricanAmerican	4208	281	47	0.89
H	AmericanIndian	4208	173	47	0.89
H	Asian	4208	125	47	0.87
H	White	4208	3386	47	0.89
H	Hispanic	4208	764	47	0.89
H	SWD	4208	236	47	0.88
H	ESOL	4208	428	47	0.89

Table D.48: Marginal Scaled Score Reliability for Grade 3 ELA Subgroups by Form A

Form	Group	N Population	N	Max Score	Reliability
A	AfricanAmerican	7812	663	74	0.93
A	AmericanIndian	7812	273	74	0.93
A	Asian	7812	180	74	0.91
A	White	7812	6174	74	0.92
A	Hispanic	7812	1757	74	0.93
A	SWD	7812	2443	74	0.93
A	ESOL	7812	1251	74	0.93

Table D.49: Marginal Scaled Score Reliability for Grade 3 ELA Subgroups by Form B

Form	Group	N Population	N	Max Score	Reliability
B	AfricanAmerican	4137	287	58	0.92
B	AmericanIndian	4137	121	58	0.92
B	Asian	4137	126	58	0.89
B	White	4137	3326	58	0.90
B	Hispanic	4137	760	58	0.91
B	SWD	4137	309	58	0.91
B	ESOL	4137	480	58	0.92

Table D.50: Marginal Scaled Score Reliability for Grade 3 ELA Subgroups by Form C

Form	Group	N Population	N	Max Score	Reliability
C	AfricanAmerican	4200	290	73	0.91
C	AmericanIndian	4200	115	73	0.91
C	Asian	4200	127	73	0.89
C	White	4200	3404	73	0.90
C	Hispanic	4200	755	73	0.91
C	SWD	4200	279	73	0.90
C	ESOL	4200	476	73	0.91

Table D.51: Marginal Scaled Score Reliability for Grade 3 ELA Subgroups by Form D

Form	Group	N Population	N	Max Score	Reliability
D	AfricanAmerican	4129	309	58	0.90
D	AmericanIndian	4129	126	58	0.90
D	Asian	4129	131	58	0.89
D	White	4129	3291	58	0.89
D	Hispanic	4129	789	58	0.90
D	SWD	4129	302	58	0.90
D	ESOL	4129	523	58	0.90

Table D.52: Marginal Scaled Score Reliability for Grade 3 ELA Subgroups by Form E

Form	Group	N Population	N	Max Score	Reliability
E	AfricanAmerican	4205	315	77	0.92
E	AmericanIndian	4205	123	77	0.92
E	Asian	4205	116	77	0.91
E	White	4205	3379	77	0.91
E	Hispanic	4205	784	77	0.92
E	SWD	4205	309	77	0.92
E	ESOL	4205	486	77	0.92

Table D.53: Marginal Scaled Score Reliability for Grade 3 ELA Subgroups by Form F

Form	Group	N Population	N	Max Score	Reliability
F	AfricanAmerican	4150	298	76	0.91
F	AmericanIndian	4150	133	76	0.91
F	Asian	4150	130	76	0.89
F	White	4150	3331	76	0.91
F	Hispanic	4150	797	76	0.91
F	SWD	4150	291	76	0.91
F	ESOL	4150	511	76	0.91

Table D.54: Marginal Scaled Score Reliability for Grade 3 ELA Subgroups by Form G

Form	Group	N Population	N	Max Score	Reliability
G	AfricanAmerican	4177	313	76	0.92
G	AmericanIndian	4177	117	76	0.92
G	Asian	4177	115	76	0.90
G	White	4177	3355	76	0.91
G	Hispanic	4177	763	76	0.92
G	SWD	4177	315	76	0.92
G	ESOL	4177	472	76	0.92

Table D.55: Marginal Scaled Score Reliability for Grade 3 ELA Subgroups by Form H

Form	Group	N Population	N	Max Score	Reliability
H	AfricanAmerican	4150	310	74	0.92
H	AmericanIndian	4150	111	74	0.92
H	Asian	4150	106	74	0.90
H	White	4150	3373	74	0.90
H	Hispanic	4150	801	74	0.91
H	SWD	4150	288	74	0.91
H	ESOL	4150	508	74	0.92

Table D.56: Marginal Scaled Score Reliability for Grade 4 ELA Subgroups by Form A

Form	Group	N Population	N	Max Score	Reliability
A	AfricanAmerican	8082	679	80	0.94
A	AmericanIndian	8082	371	80	0.94
A	Asian	8082	197	80	0.91
A	White	8082	6265	80	0.92
A	Hispanic	8082	1800	80	0.93
A	SWD	8082	2695	80	0.94
A	ESOL	8082	1327	80	0.94

Table D.57: Marginal Scaled Score Reliability for Grade 4 ELA Subgroups by Form B

Form	Group	N Population	N	Max Score	Reliability
B	AfricanAmerican	3961	271	58	0.90
B	AmericanIndian	3961	137	58	0.90
B	Asian	3961	109	58	0.86
B	White	3961	3175	58	0.88
B	Hispanic	3961	733	58	0.89
B	SWD	3961	249	58	0.89
B	ESOL	3961	439	58	0.90

Table D.58: Marginal Scaled Score Reliability for Grade 4 ELA Subgroups by Form C

Form	Group	N Population	N	Max Score	Reliability
C	AfricanAmerican	4129	277	71	0.91
C	AmericanIndian	4129	152	71	0.91
C	Asian	4129	138	71	0.88
C	White	4129	3303	71	0.89
C	Hispanic	4129	761	71	0.91
C	SWD	4129	267	71	0.91
C	ESOL	4129	523	71	0.91

Table D.59: Marginal Scaled Score Reliability for Grade 4 ELA Subgroups by Form D

Form	Group	N Population	N	Max Score	Reliability
D	AfricanAmerican	4118	280	60	0.88
D	AmericanIndian	4118	145	60	0.88
D	Asian	4118	118	60	0.85
D	White	4118	3297	60	0.86
D	Hispanic	4118	743	60	0.87
D	SWD	4118	281	60	0.88
D	ESOL	4118	489	60	0.88

Table D.60: Marginal Scaled Score Reliability for Grade 4 ELA Subgroups by Form E

Form	Group	N Population	N	Max Score	Reliability
E	AfricanAmerican	4093	285	74	0.91
E	AmericanIndian	4093	156	74	0.90
E	Asian	4093	112	74	0.88
E	White	4093	3279	74	0.89
E	Hispanic	4093	799	74	0.91
E	SWD	4093	247	74	0.91
E	ESOL	4093	468	74	0.91

Table D.61: Marginal Scaled Score Reliability for Grade 4 ELA Subgroups by Form F

Form	Group	N Population	N	Max Score	Reliability
F	AfricanAmerican	4007	296	77	0.92
F	AmericanIndian	4007	107	77	0.91
F	Asian	4007	121	77	0.88
F	White	4007	3241	77	0.90
F	Hispanic	4007	728	77	0.91
F	SWD	4007	271	77	0.91
F	ESOL	4007	478	77	0.92

Table D.62: Marginal Scaled Score Reliability for Grade 4 ELA Subgroups by Form G

Form	Group	N Population	N	Max Score	Reliability
G	AfricanAmerican	3952	299	72	0.90
G	AmericanIndian	3952	127	72	0.90
G	Asian	3952	106	72	0.88
G	White	3952	3170	72	0.89
G	Hispanic	3952	719	72	0.90
G	SWD	3952	254	72	0.90
G	ESOL	3952	437	72	0.91

Table D.63: Marginal Scaled Score Reliability for Grade 4 ELA Subgroups by Form H

Form	Group	N Population	N	Max Score	Reliability
H	AfricanAmerican	4104	251	69	0.91
H	AmericanIndian	4104	142	69	0.91
H	Asian	4104	118	69	0.88
H	White	4104	3350	69	0.89
H	Hispanic	4104	753	69	0.91
H	SWD	4104	276	69	0.90
H	ESOL	4104	482	69	0.91

Table D.64: Marginal Scaled Score Reliability for Grade 5 ELA Subgroups by Form A

Form	Group	N Population	N	Max Score	Reliability
A	AfricanAmerican	7907	634	76	0.92
A	AmericanIndian	7907	409	76	0.91
A	Asian	7907	192	76	0.88
A	White	7907	6151	76	0.90
A	Hispanic	7907	1775	76	0.91
A	SWD	7907	2718	76	0.93
A	ESOL	7907	1321	76	0.92

Table D.65: Marginal Scaled Score Reliability for Grade 5 ELA Subgroups by Form B

Form	Group	N Population	N	Max Score	Reliability
B	AfricanAmerican	3966	265	61	0.89
B	AmericanIndian	3966	130	61	0.88
B	Asian	3966	110	61	0.83
B	White	3966	3196	61	0.86
B	Hispanic	3966	693	61	0.88
B	SWD	3966	257	61	0.88
B	ESOL	3966	445	61	0.88

Table D.66: Marginal Scaled Score Reliability for Grade 5 ELA Subgroups by Form C

Form	Group	N Population	N	Max Score	Reliability
C	AfricanAmerican	3951	247	74	0.91
C	AmericanIndian	3951	136	74	0.90
C	Asian	3951	126	74	0.87
C	White	3951	3194	74	0.88
C	Hispanic	3951	758	74	0.90
C	SWD	3951	227	74	0.91
C	ESOL	3951	487	74	0.91

Table D.67: Marginal Scaled Score Reliability for Grade 5 ELA Subgroups by Form D

Form	Group	N Population	N	Max Score	Reliability
D	AfricanAmerican	4068	282	61	0.86
D	AmericanIndian	4068	156	61	0.85
D	Asian	4068	131	61	0.80
D	White	4068	3227	61	0.83
D	Hispanic	4068	739	61	0.85
D	SWD	4068	247	61	0.86
D	ESOL	4068	472	61	0.86

Table D.68: Marginal Scaled Score Reliability for Grade 5 ELA Subgroups by Form E

Form	Group	N Population	N	Max Score	Reliability
E	AfricanAmerican	4140	285	77	0.90
E	AmericanIndian	4140	148	77	0.90
E	Asian	4140	125	77	0.88
E	White	4140	3331	77	0.89
E	Hispanic	4140	745	77	0.90
E	SWD	4140	239	77	0.91
E	ESOL	4140	456	77	0.91

Table D.69: Marginal Scaled Score Reliability for Grade 5 ELA Subgroups by Form F

Form	Group	N Population	N	Max Score	Reliability
F	AfricanAmerican	3915	279	71	0.90
F	AmericanIndian	3915	150	71	0.90
F	Asian	3915	121	71	0.88
F	White	3915	3101	71	0.88
F	Hispanic	3915	679	71	0.90
F	SWD	3915	229	71	0.90
F	ESOL	3915	427	71	0.90

Table D.70: Marginal Scaled Score Reliability for Grade 5 ELA Subgroups by Form G

Form	Group	N Population	N	Max Score	Reliability
G	AfricanAmerican	4077	283	77	0.91
G	AmericanIndian	4077	174	77	0.91
G	Asian	4077	126	77	0.86
G	White	4077	3243	77	0.88
G	Hispanic	4077	744	77	0.90
G	SWD	4077	248	77	0.91
G	ESOL	4077	451	77	0.90

Table D.71: Marginal Scaled Score Reliability for Grade 5 ELA Subgroups by Form HJ

Form	Group	N Population	N	Max Score	Reliability
H	AfricanAmerican	4058	265	75	0.90
H	AmericanIndian	4058	150	75	0.90
H	Asian	4058	117	75	0.87
H	White	4058	3225	75	0.88
H	Hispanic	4058	779	75	0.90
H	SWD	4058	237	75	0.90
H	ESOL	4058	491	75	0.90

Table D.72: Marginal Scaled Score Reliability for Grade 6 ELA Subgroups by Form A

Form	Group	N Population	N	Max Score	Reliability
A	AfricanAmerican	7465	615	73	0.93
A	AmericanIndian	7465	364	73	0.92
A	Asian	7465	183	73	0.89
A	White	7465	5809	73	0.91
A	Hispanic	7465	1558	73	0.92
A	SWD	7465	2602	73	0.93
A	ESOL	7465	1092	73	0.92

Table D.73: Marginal Scaled Score Reliability for Grade 6 ELA Subgroups by Form B

Form	Group	N Population	N	Max Score	Reliability
B	AfricanAmerican	4143	294	57	0.89
B	AmericanIndian	4143	159	57	0.88
B	Asian	4143	129	57	0.85
B	White	4143	3328	57	0.86
B	Hispanic	4143	739	57	0.88
B	SWD	4143	234	57	0.89
B	ESOL	4143	456	57	0.89

Table D.74: Marginal Scaled Score Reliability for Grade 6 ELA Subgroups by Form C

Form	Group	N Population	N	Max Score	Reliability
C	AfricanAmerican	4097	258	69	0.91
C	AmericanIndian	4097	167	69	0.91
C	Asian	4097	112	69	0.88
C	White	4097	3306	69	0.89
C	Hispanic	4097	745	69	0.91
C	SWD	4097	211	69	0.91
C	ESOL	4097	475	69	0.91

Table D.75: Marginal Scaled Score Reliability for Grade 6 ELA Subgroups by Form D

Form	Group	N Population	N	Max Score	Reliability
D	AfricanAmerican	4027	283	58	0.89
D	AmericanIndian	4027	176	58	0.88
D	Asian	4027	121	58	0.85
D	White	4027	3237	58	0.86
D	Hispanic	4027	741	58	0.88
D	SWD	4027	244	58	0.88
D	ESOL	4027	459	58	0.89

Table D.76: Marginal Scaled Score Reliability for Grade 6 ELA Subgroups by Form E

Form	Group	N Population	N	Max Score	Reliability
E	AfricanAmerican	4242	324	73	0.90
E	AmericanIndian	4242	175	73	0.90
E	Asian	4242	135	73	0.88
E	White	4242	3357	73	0.89
E	Hispanic	4242	740	73	0.90
E	SWD	4242	265	73	0.90
E	ESOL	4242	483	73	0.91

Table D.77: Marginal Scaled Score Reliability for Grade 6 ELA Subgroups by Form F

Form	Group	N Population	N	Max Score	Reliability
F	AfricanAmerican	4053	293	72	0.90
F	AmericanIndian	4053	166	72	0.90
F	Asian	4053	127	72	0.88
F	White	4053	3238	72	0.89
F	Hispanic	4053	714	72	0.90
F	SWD	4053	216	72	0.90
F	ESOL	4053	461	72	0.90

Table D.78: Marginal Scaled Score Reliability for Grade 6 ELA Subgroups by Form G

Form	Group	N Population	N	Max Score	Reliability
G	AfricanAmerican	4130	303	69	0.91
G	AmericanIndian	4130	167	69	0.91
G	Asian	4130	115	69	0.88
G	White	4130	3297	69	0.89
G	Hispanic	4130	758	69	0.91
G	SWD	4130	246	69	0.91
G	ESOL	4130	484	69	0.91

Table D.79: Marginal Scaled Score Reliability for Grade 6 ELA Subgroups by Form H

Form	Group	N Population	N	Max Score	Reliability
H	AfricanAmerican	4253	301	66	0.90
H	AmericanIndian	4253	182	66	0.90
H	Asian	4253	127	66	0.87
H	White	4253	3397	66	0.88
H	Hispanic	4253	791	66	0.90
H	SWD	4253	241	66	0.90
H	ESOL	4253	505	66	0.90

Table D.80: Marginal Scaled Score Reliability for Grade 7 ELA Subgroups by Form A

Form	Group	N Population	N	Max Score	Reliability
A	AfricanAmerican	7288	567	72	0.92
A	AmericanIndian	7288	372	72	0.92
A	Asian	7288	174	72	0.90
A	White	7288	5777	72	0.91
A	Hispanic	7288	1513	72	0.92
A	SWD	7288	2558	72	0.93
A	ESOL	7288	1056	72	0.92

Table D.81: Marginal Scaled Score Reliability for Grade 7 ELA Subgroups by Form B

Form	Group	N Population	N	Max Score	Reliability
B	AfricanAmerican	4134	313	58	0.88
B	AmericanIndian	4134	163	58	0.88
B	Asian	4134	126	58	0.86
B	White	4134	3255	58	0.87
B	Hispanic	4134	729	58	0.88
B	SWD	4134	258	58	0.88
B	ESOL	4134	446	58	0.88

Table D.82: Marginal Scaled Score Reliability for Grade 7 ELA Subgroups by Form C

Form	Group	N Population	N	Max Score	Reliability
C	AfricanAmerican	4030	309	74	0.91
C	AmericanIndian	4030	165	74	0.91
C	Asian	4030	109	74	0.88
C	White	4030	3187	74	0.89
C	Hispanic	4030	715	74	0.91
C	SWD	4030	230	74	0.92
C	ESOL	4030	433	74	0.91

Table D.83: Marginal Scaled Score Reliability for Grade 7 ELA Subgroups by Form D

Form	Group	N Population	N	Max Score	Reliability
D	AfricanAmerican	4158	307	58	0.85
D	AmericanIndian	4158	179	58	0.85
D	Asian	4158	117	58	0.82
D	White	4158	3322	58	0.83
D	Hispanic	4158	699	58	0.84
D	SWD	4158	204	58	0.86
D	ESOL	4158	443	58	0.85

Table D.84: Marginal Scaled Score Reliability for Grade 7 ELA Subgroups by Form E

Form	Group	N Population	N	Max Score	Reliability
E	AfricanAmerican	4171	344	80	0.90
E	AmericanIndian	4171	157	80	0.90
E	Asian	4171	139	80	0.87
E	White	4171	3309	80	0.89
E	Hispanic	4171	717	80	0.90
E	SWD	4171	239	80	0.91
E	ESOL	4171	452	80	0.90

Table D.85: Marginal Scaled Score Reliability for Grade 7 ELA Subgroups by Form F

Form	Group	N Population	N	Max Score	Reliability
F	AfricanAmerican	4124	280	82	0.89
F	AmericanIndian	4124	190	82	0.89
F	Asian	4124	125	82	0.87
F	White	4124	3270	82	0.87
F	Hispanic	4124	730	82	0.89
F	SWD	4124	226	82	0.89
F	ESOL	4124	440	82	0.89

Table D.86: Marginal Scaled Score Reliability for Grade 7 ELA Subgroups by Form G

Form	Group	N Population	N	Max Score	Reliability
G	AfricanAmerican	4119	277	80	0.89
G	AmericanIndian	4119	170	80	0.90
G	Asian	4119	117	80	0.87
G	White	4119	3288	80	0.88
G	Hispanic	4119	731	80	0.89
G	SWD	4119	216	80	0.90
G	ESOL	4119	446	80	0.90

Table D.87: Marginal Scaled Score Reliability for Grade 7 ELA Subgroups by Form H

Form	Group	N Population	N	Max Score	Reliability
H	AfricanAmerican	4009	287	76	0.89
H	AmericanIndian	4009	166	76	0.89
H	Asian	4009	112	76	0.86
H	White	4009	3203	76	0.87
H	Hispanic	4009	670	76	0.89
H	SWD	4009	216	76	0.90
H	ESOL	4009	422	76	0.89

Table D.88: Marginal Scaled Score Reliability for Grade 8 ELA Subgroups by Form A

Form	Group	N Population	N	Max Score	Reliability
A	AfricanAmerican	6718	535	75	0.93
A	AmericanIndian	6718	348	75	0.93
A	Asian	6718	160	75	0.89
A	White	6718	5297	75	0.91
A	Hispanic	6718	1400	75	0.93
A	SWD	6718	2313	75	0.93
A	ESOL	6718	919	75	0.93

Table D.89: Marginal Scaled Score Reliability for Grade 8 ELA Subgroups by Form B

Form	Group	N Population	N	Max Score	Reliability
B	AfricanAmerican	4107	262	57	0.89
B	AmericanIndian	4107	165	57	0.90
B	Asian	4107	130	57	0.86
B	White	4107	3311	57	0.87
B	Hispanic	4107	727	57	0.89
B	SWD	4107	221	57	0.90
B	ESOL	4107	422	57	0.90

Table D.90: Marginal Scaled Score Reliability for Grade 8 ELA Subgroups by Form C

Form	Group	N Population	N	Max Score	Reliability
C	AfricanAmerican	4198	315	68	0.91
C	AmericanIndian	4198	160	68	0.91
C	Asian	4198	102	68	0.88
C	White	4198	3359	68	0.89
C	Hispanic	4198	691	68	0.91
C	SWD	4198	214	68	0.91
C	ESOL	4198	387	68	0.91

Table D.91: Marginal Scaled Score Reliability for Grade 8 ELA Subgroups by Form D

Form	Group	N Population	N	Max Score	Reliability
D	AfricanAmerican	4161	270	55	0.89
D	AmericanIndian	4161	173	55	0.88
D	Asian	4161	124	55	0.85
D	White	4161	3345	55	0.87
D	Hispanic	4161	759	55	0.88
D	SWD	4161	253	55	0.89
D	ESOL	4161	428	55	0.88

Table D.92: Marginal Scaled Score Reliability for Grade 8 ELA Subgroups by Form E

Form	Group	N Population	N	Max Score	Reliability
E	AfricanAmerican	4218	291	64	0.89
E	AmericanIndian	4218	173	64	0.89
E	Asian	4218	125	64	0.86
E	White	4218	3385	64	0.87
E	Hispanic	4218	747	64	0.88
E	SWD	4218	200	64	0.89
E	ESOL	4218	421	64	0.89

Table D.93: Marginal Scaled Score Reliability for Grade 8 ELA Subgroups by Form F

Form	Group	N Population	N	Max Score	Reliability
F	AfricanAmerican	4133	297	76	0.91
F	AmericanIndian	4133	162	76	0.91
F	Asian	4133	107	76	0.88
F	White	4133	3345	76	0.89
F	Hispanic	4133	730	76	0.90
F	SWD	4133	231	76	0.92
F	ESOL	4133	408	76	0.91

Table D.94: Marginal Scaled Score Reliability for Grade 8 ELA Subgroups by Form G

Form	Group	N Population	N	Max Score	Reliability
G	AfricanAmerican	4149	278	73	0.91
G	AmericanIndian	4149	179	73	0.91
G	Asian	4149	122	73	0.89
G	White	4149	3316	73	0.90
G	Hispanic	4149	733	73	0.91
G	SWD	4149	243	73	0.92
G	ESOL	4149	429	73	0.91

Table D.95: Marginal Scaled Score Reliability for Grade 8 ELA Subgroups by Form H

Form	Group	N Population	N	Max Score	Reliability
H	AfricanAmerican	4113	278	70	0.90
H	AmericanIndian	4113	176	70	0.90
H	Asian	4113	125	70	0.87
H	White	4113	3274	70	0.88
H	Hispanic	4113	694	70	0.90
H	SWD	4113	208	70	0.90
H	ESOL	4113	403	70	0.90

Table D.96: Marginal Scaled Score Reliability for Grade 10 ELA Subgroups by Form A

Form	Group	N Population	N	Max Score	Reliability
A	AfricanAmerican	5650	383	71	0.92
A	AmericanIndian	5650	265	71	0.91
A	Asian	5650	116	71	0.90
A	White	5650	4557	71	0.90
A	Hispanic	5650	894	71	0.91
A	SWD	5650	1568	71	0.92
A	ESOL	5650	497	71	0.92

Table D.97: Marginal Scaled Score Reliability for Grade 10 ELA Subgroups by Form B

Form	Group	N Population	N	Max Score	Reliability
B	AfricanAmerican	4167	284	74	0.91
B	AmericanIndian	4167	175	74	0.91
B	Asian	4167	124	74	0.90
B	White	4167	3336	74	0.90
B	Hispanic	4167	682	74	0.91
B	SWD	4167	258	74	0.92
B	ESOL	4167	371	74	0.92

Table D.98: Marginal Scaled Score Reliability for Grade 10 ELA Subgroups by Form C

Form	Group	N Population	N	Max Score	Reliability
C	AfricanAmerican	4202	277	72	0.90
C	AmericanIndian	4202	177	72	0.90
C	Asian	4202	125	72	0.88
C	White	4202	3366	72	0.88
C	Hispanic	4202	739	72	0.90
C	SWD	4202	252	72	0.91
C	ESOL	4202	363	72	0.91

Table D.99: Marginal Scaled Score Reliability for Grade 10 ELA Subgroups by Form D

Form	Group	N Population	N	Max Score	Reliability
D	AfricanAmerican	4159	272	85	0.92
D	AmericanIndian	4159	181	85	0.91
D	Asian	4159	106	85	0.90
D	White	4159	3343	85	0.90
D	Hispanic	4159	728	85	0.91
D	SWD	4159	271	85	0.92
D	ESOL	4159	367	85	0.92

Table D.100: Marginal Scaled Score Reliability for Grade 10 ELA Subgroups by Form E

Form	Group	N Population	N	Max Score	Reliability
E	AfricanAmerican	4218	266	71	0.90
E	AmericanIndian	4218	171	71	0.90
E	Asian	4218	119	71	0.88
E	White	4218	3430	71	0.88
E	Hispanic	4218	700	71	0.90
E	SWD	4218	283	71	0.91
E	ESOL	4218	346	71	0.91

Table D.101: Marginal Scaled Score Reliability for Grade 10 ELA Subgroups by Form F

Form	Group	N Population	N	Max Score	Reliability
F	AfricanAmerican	4148	286	82	0.91
F	AmericanIndian	4148	153	82	0.90
F	Asian	4148	112	82	0.88
F	White	4148	3341	82	0.89
F	Hispanic	4148	716	82	0.90
F	SWD	4148	274	82	0.91
F	ESOL	4148	359	82	0.91

Table D.102: Marginal Scaled Score Reliability for Grade 10 ELA Subgroups by Form G

Form	Group	N Population	N	Max Score	Reliability
G	AfricanAmerican	4171	279	83	0.91
G	AmericanIndian	4171	165	83	0.91
G	Asian	4171	125	83	0.90
G	White	4171	3364	83	0.91
G	Hispanic	4171	699	83	0.91
G	SWD	4171	268	83	0.91
G	ESOL	4171	374	83	0.91

Table D.103: Marginal Scaled Score Reliability for Grade 10 ELA Subgroups by Form H

Form	Group	N Population	N	Max Score	Reliability
H	AfricanAmerican	4135	273	83	0.92
H	AmericanIndian	4135	179	83	0.92
H	Asian	4135	102	83	0.90
H	White	4135	3327	83	0.90
H	Hispanic	4135	691	83	0.91
H	SWD	4135	275	83	0.92
H	ESOL	4135	342	83	0.92

E
Claim-Score Reliability

Table E.1: Marginal Scaled Score Reliability for Grade 3 Math Claim Scores

Subject	Grade	Form	Claim	Max Score	Reliability
Math	3	A	Claim 1	44	0.88
Math	3	A	Claim 2	8	0.56
Math	3	A	Claim 3	6	0.51
Math	3	A	Claim 4	6	0.42
Math	3	A	Claims 2, 3 and 4	20	0.75
Math	3	B	Claim 1	36	0.87
Math	3	B	Claim 2	5	0.39
Math	3	B	Claim 3	6	0.49
Math	3	B	Claim 4	4	0.40
Math	3	B	Claims 2, 3 and 4	15	0.69
Math	3	C	Claim 1	44	0.90
Math	3	C	Claim 2	8	0.46
Math	3	C	Claim 3	4	0.45
Math	3	C	Claim 4	7	0.40
Math	3	C	Claims 2, 3 and 4	19	0.68
Math	3	D	Claim 1	45	0.90
Math	3	D	Claim 2	7	0.52
Math	3	D	Claim 3	7	0.57
Math	3	D	Claim 4	7	0.52
Math	3	D	Claims 2, 3 and 4	21	0.77
Math	3	E	Claim 1	36	0.88
Math	3	E	Claim 2	6	0.49
Math	3	E	Claim 3	4	0.44
Math	3	E	Claim 4	6	0.44
Math	3	E	Claims 2, 3 and 4	16	0.71
Math	3	F	Claim 1	44	0.90
Math	3	F	Claim 2	6	0.42
Math	3	F	Claim 3	5	0.52
Math	3	F	Claim 4	6	0.47
Math	3	F	Claims 2, 3 and 4	17	0.72
Math	3	G	Claim 1	45	0.89
Math	3	G	Claim 2	7	0.42
Math	3	G	Claim 3	6	0.54
Math	3	G	Claim 4	6	0.46
Math	3	G	Claims 2, 3 and 4	19	0.73
Math	3	H	Claim 1	44	0.90
Math	3	H	Claim 2	8	0.57
Math	3	H	Claim 3	5	0.47
Math	3	H	Claim 4	6	0.48
Math	3	H	Claims 2, 3 and 4	19	0.75

Table E.2: Marginal Scaled Score Reliability for Grade 4 Math Claim Scores

Subject	Grade	Form	Claim	Max Score	Reliability
Math	4	A	Claim 1	45	0.90
Math	4	A	Claim 2	7	0.50
Math	4	A	Claim 3	7	0.46
Math	4	A	Claim 4	8	0.51
Math	4	A	Claims 2, 3 and 4	22	0.73
Math	4	B	Claim 1	34	0.87
Math	4	B	Claim 2	3	0.19
Math	4	B	Claim 3	4	0.47
Math	4	B	Claim 4	5	0.52
Math	4	B	Claims 2, 3 and 4	12	0.68
Math	4	C	Claim 1	46	0.90
Math	4	C	Claim 2	7	0.42
Math	4	C	Claim 3	7	0.49
Math	4	C	Claim 4	7	0.41
Math	4	C	Claims 2, 3 and 4	21	0.70
Math	4	D	Claim 1	45	0.90
Math	4	D	Claim 2	7	0.51
Math	4	D	Claim 3	9	0.65
Math	4	D	Claim 4	5	0.45
Math	4	D	Claims 2, 3 and 4	21	0.78
Math	4	E	Claim 1	36	0.89
Math	4	E	Claim 2	6	0.38
Math	4	E	Claim 3	4	0.56
Math	4	E	Claim 4	5	0.36
Math	4	E	Claims 2, 3 and 4	15	0.68
Math	4	F	Claim 1	46	0.91
Math	4	F	Claim 2	7	0.50
Math	4	F	Claim 3	7	0.53
Math	4	F	Claim 4	8	0.53
Math	4	F	Claims 2, 3 and 4	22	0.76
Math	4	G	Claim 1	46	0.90
Math	4	G	Claim 2	7	0.49
Math	4	G	Claim 3	9	0.65
Math	4	G	Claim 4	5	0.45
Math	4	G	Claims 2, 3 and 4	21	0.78
Math	4	H	Claim 1	44	0.90
Math	4	H	Claim 2	6	0.48
Math	4	H	Claim 3	7	0.58
Math	4	H	Claim 4	7	0.60
Math	4	H	Claims 2, 3 and 4	20	0.78

Table E.3: Marginal Scaled Score Reliability for Grade 5 Math Claim Scores

Subject	Grade	Form	Claim	Max Score	Reliability
Math	5	A	Claim 1	45	0.90
Math	5	A	Claim 2	7	0.56
Math	5	A	Claim 3	8	0.62
Math	5	A	Claim 4	8	0.51
Math	5	A	Claims 2, 3 and 4	23	0.79
Math	5	B	Claim 1	36	0.88
Math	5	B	Claim 2	5	0.53
Math	5	B	Claim 3	5	0.40
Math	5	B	Claim 4	6	0.41
Math	5	B	Claims 2, 3 and 4	16	0.70
Math	5	C	Claim 1	44	0.90
Math	5	C	Claim 2	6	0.43
Math	5	C	Claim 3	8	0.60
Math	5	C	Claim 4	7	0.52
Math	5	C	Claims 2, 3 and 4	21	0.77
Math	5	D	Claim 1	45	0.91
Math	5	D	Claim 2	6	0.33
Math	5	D	Claim 3	7	0.50
Math	5	D	Claim 4	7	0.45
Math	5	D	Claims 2, 3 and 4	20	0.69
Math	5	E	Claim 1	37	0.89
Math	5	E	Claim 2	4	0.29
Math	5	E	Claim 3	5	0.43
Math	5	E	Claim 4	5	0.35
Math	5	E	Claims 2, 3 and 4	14	0.61
Math	5	F	Claim 1	44	0.90
Math	5	F	Claim 2	5	0.34
Math	5	F	Claim 3	6	0.53
Math	5	F	Claim 4	7	0.42
Math	5	F	Claims 2, 3 and 4	18	0.69
Math	5	G	Claim 1	44	0.91
Math	5	G	Claim 2	7	0.44
Math	5	G	Claim 3	7	0.50
Math	5	G	Claim 4	6	0.48
Math	5	G	Claims 2, 3 and 4	20	0.72
Math	5	H	Claim 1	44	0.90
Math	5	H	Claim 2	8	0.54
Math	5	H	Claim 3	7	0.48
Math	5	H	Claim 4	6	0.49
Math	5	H	Claims 2, 3 and 4	21	0.74

Table E.4: Marginal Scaled Score Reliability for Grade 6 Math Claim Scores

Subject	Grade	Form	Claim	Max Score	Reliability
Math	6	A	Claim 1	46	0.89
Math	6	A	Claim 2	7	0.28
Math	6	A	Claim 3	8	0.36
Math	6	A	Claim 4	8	0.32
Math	6	A	Claims 2, 3 and 4	23	0.59
Math	6	B	Claim 1	34	0.85
Math	6	B	Claim 2	6	0.45
Math	6	B	Claim 3	5	0.32
Math	6	B	Claim 4	5	0.23
Math	6	B	Claims 2, 3 and 4	16	0.60
Math	6	C	Claim 1	43	0.89
Math	6	C	Claim 2	7	0.38
Math	6	C	Claim 3	7	0.37
Math	6	C	Claim 4	7	0.31
Math	6	C	Claims 2, 3 and 4	21	0.61
Math	6	D	Claim 1	46	0.90
Math	6	D	Claim 2	6	0.42
Math	6	D	Claim 3	6	0.37
Math	6	D	Claim 4	6	0.40
Math	6	D	Claims 2, 3 and 4	18	0.66
Math	6	E	Claim 1	37	0.88
Math	6	E	Claim 2	4	0.31
Math	6	E	Claim 3	5	0.38
Math	6	E	Claim 4	3	0.28
Math	6	E	Claims 2, 3 and 4	12	0.59
Math	6	F	Claim 1	41	0.87
Math	6	F	Claim 2	6	0.38
Math	6	F	Claim 3	6	0.25
Math	6	F	Claim 4	5	0.30
Math	6	F	Claims 2, 3 and 4	17	0.57
Math	6	G	Claim 1	44	0.90
Math	6	G	Claim 2	6	0.48
Math	6	G	Claim 3	8	0.45
Math	6	G	Claim 4	7	0.37
Math	6	G	Claims 2, 3 and 4	21	0.69
Math	6	H	Claim 1	37	0.87
Math	6	H	Claim 2	5	0.39
Math	6	H	Claim 3	6	0.42
Math	6	H	Claim 4	5	0.29
Math	6	H	Claims 2, 3 and 4	16	0.63

Table E.5: Marginal Scaled Score Reliability for Grade 7 Math Claim Scores

Subject	Grade	Form	Claim	Max Score	Reliability
Math	7	A	Claim 1	42	0.87
Math	7	A	Claim 2	7	0.49
Math	7	A	Claim 3	7	0.41
Math	7	A	Claim 4	4	0.49
Math	7	A	Claims 2, 3 and 4	18	0.72
Math	7	B	Claim 1	35	0.85
Math	7	B	Claim 2	5	0.36
Math	7	B	Claim 3	5	0.33
Math	7	B	Claim 4	2	0.40
Math	7	B	Claims 2, 3 and 4	12	0.61
Math	7	C	Claim 1	44	0.84
Math	7	C	Claim 2	8	0.47
Math	7	C	Claim 3	7	0.47
Math	7	C	Claim 4	4	0.51
Math	7	C	Claims 2, 3 and 4	19	0.72
Math	7	D	Claim 1	44	0.87
Math	7	D	Claim 2	8	0.53
Math	7	D	Claim 3	6	0.38
Math	7	D	Claim 4	4	0.46
Math	7	D	Claims 2, 3 and 4	18	0.70
Math	7	E	Claim 1	35	0.83
Math	7	E	Claim 2	6	0.53
Math	7	E	Claim 3	5	0.33
Math	7	E	Claim 4	5	0.53
Math	7	E	Claims 2, 3 and 4	16	0.73
Math	7	F	Claim 1	45	0.87
Math	7	F	Claim 2	8	0.51
Math	7	F	Claim 3	6	0.43
Math	7	F	Claim 4	6	0.48
Math	7	F	Claims 2, 3 and 4	20	0.73
Math	7	G	Claim 1	42	0.86
Math	7	G	Claim 2	7	0.52
Math	7	G	Claim 3	6	0.42
Math	7	G	Claim 4	6	0.46
Math	7	G	Claims 2, 3 and 4	19	0.72
Math	7	H	Claim 1	32	0.84
Math	7	H	Claim 2	6	0.53
Math	7	H	Claim 3	5	0.42
Math	7	H	Claim 4	5	0.54
Math	7	H	Claims 2, 3 and 4	16	0.74

Table E.6: Marginal Scaled Score Reliability for Grade 8 Math Claim Scores

Subject	Grade	Form	Claim	Max Score	Reliability
Math	8	A	Claim 1	46	0.88
Math	8	A	Claim 2	6	0.20
Math	8	A	Claim 3	5	0.27
Math	8	A	Claim 4	6	0.41
Math	8	A	Claims 2, 3 and 4	17	0.55
Math	8	B	Claim 1	35	0.85
Math	8	B	Claim 2	5	0.15
Math	8	B	Claim 3	4	0.25
Math	8	B	Claim 4	5	0.46
Math	8	B	Claims 2, 3 and 4	14	0.55
Math	8	C	Claim 1	40	0.86
Math	8	C	Claim 2	4	0.20
Math	8	C	Claim 3	6	0.33
Math	8	C	Claim 4	6	0.45
Math	8	C	Claims 2, 3 and 4	16	0.60
Math	8	D	Claim 1	46	0.87
Math	8	D	Claim 2	4	0.28
Math	8	D	Claim 3	5	0.35
Math	8	D	Claim 4	6	0.50
Math	8	D	Claims 2, 3 and 4	15	0.65
Math	8	E	Claim 1	36	0.84
Math	8	E	Claim 2	4	0.20
Math	8	E	Claim 3	5	0.30
Math	8	E	Claim 4	3	0.37
Math	8	E	Claims 2, 3 and 4	12	0.52
Math	8	F	Claim 1	44	0.88
Math	8	F	Claim 2	6	0.21
Math	8	F	Claim 3	5	0.31
Math	8	F	Claim 4	4	0.42
Math	8	F	Claims 2, 3 and 4	15	0.54
Math	8	G	Claim 1	44	0.87
Math	8	G	Claim 2	4	0.26
Math	8	G	Claim 3	7	0.34
Math	8	G	Claim 4	6	0.48
Math	8	G	Claims 2, 3 and 4	17	0.63
Math	8	H	Claim 1	35	0.87
Math	8	H	Claim 2	3	0.15
Math	8	H	Claim 3	5	0.30
Math	8	H	Claim 4	4	0.39
Math	8	H	Claims 2, 3 and 4	12	0.53

Table E.7: Marginal Scaled Score Reliability for Grade 10 Math Claim Scores

Subject	Grade	Form	Claim	Max Score	Reliability
Math	10	A	Claim 1	42	0.87
Math	10	A	Claim 2	5	0.32
Math	10	A	Claim 3	6	0.30
Math	10	A	Claim 4	4	0.16
Math	10	A	Claims 2, 3 and 4	15	0.52
Math	10	B	Claim 1	37	0.87
Math	10	B	Claim 2	8	0.45
Math	10	B	Claim 3	5	0.26
Math	10	B	Claim 4	5	0.22
Math	10	B	Claims 2, 3 and 4	18	0.58
Math	10	C	Claim 1	41	0.87
Math	10	C	Claim 2	8	0.42
Math	10	C	Claim 3	4	0.38
Math	10	C	Claim 4	6	0.21
Math	10	C	Claims 2, 3 and 4	18	0.60
Math	10	D	Claim 1	43	0.86
Math	10	D	Claim 2	6	0.34
Math	10	D	Claim 3	6	0.30
Math	10	D	Claim 4	3	0.13
Math	10	D	Claims 2, 3 and 4	15	0.52
Math	10	E	Claim 1	42	0.88
Math	10	E	Claim 2	5	0.33
Math	10	E	Claim 3	4	0.36
Math	10	E	Claim 4	6	0.17
Math	10	E	Claims 2, 3 and 4	15	0.54
Math	10	F	Claim 1	40	0.87
Math	10	F	Claim 2	6	0.33
Math	10	F	Claim 3	5	0.30
Math	10	F	Claim 4	5	0.18
Math	10	F	Claims 2, 3 and 4	16	0.53
Math	10	G	Claim 1	45	0.87
Math	10	G	Claim 2	5	0.32
Math	10	G	Claim 3	5	0.27
Math	10	G	Claim 4	5	0.20
Math	10	G	Claims 2, 3 and 4	15	0.50
Math	10	H	Claim 1	35	0.87
Math	10	H	Claim 2	7	0.39
Math	10	H	Claim 3	4	0.33
Math	10	H	Claim 4	6	0.18
Math	10	H	Claims 2, 3 and 4	17	0.55

Table E.8: Marginal Scaled Score Reliability for Grade 3 ELA Claim Scores

Subject	Grade	Form	Claim	Max Score	Reliability
ELA	3	A	Claim 1	45	0.86
ELA	3	A	Claim 1 Information	21	0.78
ELA	3	A	Claim 1 Literary	24	0.73
ELA	3	A	Claim 2	29	0.80
ELA	3	B	Claim 1	39	0.87
ELA	3	B	Claim 1 Information	21	0.78
ELA	3	B	Claim 1 Literary	18	0.75
ELA	3	B	Claim 2	19	0.74
ELA	3	C	Claim 1	45	0.85
ELA	3	C	Claim 1 Information	20	0.74
ELA	3	C	Claim 1 Literary	25	0.74
ELA	3	C	Claim 2	28	0.77
ELA	3	D	Claim 1	38	0.85
ELA	3	D	Claim 1 Information	21	0.78
ELA	3	D	Claim 1 Literary	17	0.70
ELA	3	D	Claim 2	20	0.72
ELA	3	E	Claim 1	50	0.87
ELA	3	E	Claim 1 Information	22	0.79
ELA	3	E	Claim 1 Literary	28	0.75
ELA	3	E	Claim 2	27	0.80
ELA	3	F	Claim 1	46	0.86
ELA	3	F	Claim 1 Information	20	0.73
ELA	3	F	Claim 1 Literary	26	0.78
ELA	3	F	Claim 2	30	0.78
ELA	3	G	Claim 1	49	0.88
ELA	3	G	Claim 1 Information	31	0.82
ELA	3	G	Claim 1 Literary	18	0.75
ELA	3	G	Claim 2	27	0.75
ELA	3	H	Claim 1	48	0.86
ELA	3	H	Claim 1 Information	21	0.73
ELA	3	H	Claim 1 Literary	27	0.78
ELA	3	H	Claim 2	26	0.77

Table E.9: Marginal Scaled Score Reliability for Grade 4 ELA Claim Scores

Subject	Grade	Form	Claim	Max Score	Reliability
ELA	4	A	Claim 1	51	0.87
ELA	4	A	Claim 1 Information	20	0.74
ELA	4	A	Claim 1 Literary	31	0.80
ELA	4	A	Claim 2	29	0.79
ELA	4	B	Claim 1	41	0.83
ELA	4	B	Claim 1 Information	21	0.74
ELA	4	B	Claim 1 Literary	20	0.69
ELA	4	B	Claim 2	17	0.72
ELA	4	C	Claim 1	49	0.85
ELA	4	C	Claim 1 Information	29	0.80
ELA	4	C	Claim 1 Literary	20	0.67
ELA	4	C	Claim 2	22	0.71
ELA	4	D	Claim 1	40	0.80
ELA	4	D	Claim 1 Information	20	0.71
ELA	4	D	Claim 1 Literary	20	0.64
ELA	4	D	Claim 2	20	0.69
ELA	4	E	Claim 1	46	0.85
ELA	4	E	Claim 1 Information	27	0.80
ELA	4	E	Claim 1 Literary	19	0.67
ELA	4	E	Claim 2	28	0.75
ELA	4	F	Claim 1	49	0.85
ELA	4	F	Claim 1 Information	20	0.74
ELA	4	F	Claim 1 Literary	29	0.75
ELA	4	F	Claim 2	28	0.77
ELA	4	G	Claim 1	48	0.85
ELA	4	G	Claim 1 Information	30	0.80
ELA	4	G	Claim 1 Literary	18	0.68
ELA	4	G	Claim 2	24	0.71
ELA	4	H	Claim 1	47	0.86
ELA	4	H	Claim 1 Information	20	0.75
ELA	4	H	Claim 1 Literary	27	0.77
ELA	4	H	Claim 2	22	0.73

Table E.10: Marginal Scaled Score Reliability for Grade 5 ELA Claim Scores

Subject	Grade	Form	Claim	Max Score	Reliability
ELA	5	A	Claim 1	48	0.81
ELA	5	A	Claim 1 Information	20	0.69
ELA	5	A	Claim 1 Literary	28	0.69
ELA	5	A	Claim 2	28	0.75
ELA	5	B	Claim 1	40	0.80
ELA	5	B	Claim 1 Information	21	0.71
ELA	5	B	Claim 1 Literary	19	0.63
ELA	5	B	Claim 2	21	0.69
ELA	5	C	Claim 1	47	0.83
ELA	5	C	Claim 1 Information	32	0.79
ELA	5	C	Claim 1 Literary	15	0.57
ELA	5	C	Claim 2	27	0.75
ELA	5	D	Claim 1	39	0.77
ELA	5	D	Claim 1 Information	20	0.65
ELA	5	D	Claim 1 Literary	19	0.62
ELA	5	D	Claim 2	22	0.65
ELA	5	E	Claim 1	48	0.83
ELA	5	E	Claim 1 Information	29	0.76
ELA	5	E	Claim 1 Literary	19	0.63
ELA	5	E	Claim 2	29	0.77
ELA	5	F	Claim 1	46	0.84
ELA	5	F	Claim 1 Information	29	0.77
ELA	5	F	Claim 1 Literary	17	0.68
ELA	5	F	Claim 2	25	0.72
ELA	5	G	Claim 1	46	0.83
ELA	5	G	Claim 1 Information	20	0.72
ELA	5	G	Claim 1 Literary	26	0.71
ELA	5	G	Claim 2	31	0.75
ELA	5	H	Claim 1	48	0.84
ELA	5	H	Claim 1 Information	29	0.77
ELA	5	H	Claim 1 Literary	19	0.68
ELA	5	H	Claim 2	27	0.72

Table E.11: Marginal Scaled Score Reliability for Grade 6 ELA Claim Scores

Subject	Grade	Form	Claim	Max Score	Reliability
ELA	6	A	Claim 1	47	0.86
ELA	6	A	Claim 1 Information	20	0.74
ELA	6	A	Claim 1 Literary	27	0.77
ELA	6	A	Claim 2	26	0.76
ELA	6	B	Claim 1	36	0.81
ELA	6	B	Claim 1 Information	19	0.69
ELA	6	B	Claim 1 Literary	17	0.68
ELA	6	B	Claim 2	21	0.70
ELA	6	C	Claim 1	45	0.86
ELA	6	C	Claim 1 Information	29	0.80
ELA	6	C	Claim 1 Literary	16	0.69
ELA	6	C	Claim 2	24	0.73
ELA	6	D	Claim 1	37	0.83
ELA	6	D	Claim 1 Information	21	0.71
ELA	6	D	Claim 1 Literary	16	0.72
ELA	6	D	Claim 2	21	0.64
ELA	6	E	Claim 1	43	0.84
ELA	6	E	Claim 1 Information	18	0.73
ELA	6	E	Claim 1 Literary	25	0.74
ELA	6	E	Claim 2	30	0.75
ELA	6	F	Claim 1	46	0.85
ELA	6	F	Claim 1 Information	21	0.74
ELA	6	F	Claim 1 Literary	25	0.74
ELA	6	F	Claim 2	26	0.72
ELA	6	G	Claim 1	47	0.86
ELA	6	G	Claim 1 Information	19	0.76
ELA	6	G	Claim 1 Literary	28	0.78
ELA	6	G	Claim 2	22	0.68
ELA	6	H	Claim 1	42	0.85
ELA	6	H	Claim 1 Information	26	0.79
ELA	6	H	Claim 1 Literary	16	0.69
ELA	6	H	Claim 2	24	0.69

Table E.12: Marginal Scaled Score Reliability for Grade 7 ELA Claim Scores

Subject	Grade	Form	Claim	Max Score	Reliability
ELA	7	A	Claim 1	49	0.87
ELA	7	A	Claim 1 Information	20	0.76
ELA	7	A	Claim 1 Literary	29	0.78
ELA	7	A	Claim 2	23	0.72
ELA	7	B	Claim 1	39	0.82
ELA	7	B	Claim 1 Information	22	0.73
ELA	7	B	Claim 1 Literary	17	0.66
ELA	7	B	Claim 2	19	0.68
ELA	7	C	Claim 1	49	0.85
ELA	7	C	Claim 1 Information	29	0.78
ELA	7	C	Claim 1 Literary	20	0.71
ELA	7	C	Claim 2	25	0.76
ELA	7	D	Claim 1	35	0.76
ELA	7	D	Claim 1 Information	21	0.68
ELA	7	D	Claim 1 Literary	14	0.55
ELA	7	D	Claim 2	23	0.65
ELA	7	E	Claim 1	49	0.84
ELA	7	E	Claim 1 Information	19	0.70
ELA	7	E	Claim 1 Literary	30	0.75
ELA	7	E	Claim 2	31	0.75
ELA	7	F	Claim 1	48	0.81
ELA	7	F	Claim 1 Information	29	0.72
ELA	7	F	Claim 1 Literary	19	0.64
ELA	7	F	Claim 2	34	0.75
ELA	7	G	Claim 1	47	0.83
ELA	7	G	Claim 1 Information	18	0.70
ELA	7	G	Claim 1 Literary	29	0.73
ELA	7	G	Claim 2	33	0.74
ELA	7	H	Claim 1	45	0.80
ELA	7	H	Claim 1 Information	29	0.72
ELA	7	H	Claim 1 Literary	16	0.60
ELA	7	H	Claim 2	31	0.76

Table E.13: Marginal Scaled Score Reliability for Grade 8 ELA Claim Scores

Subject	Grade	Form	Claim	Max Score	Reliability
ELA	8	A	Claim 1	47	0.87
ELA	8	A	Claim 1 Information	19	0.73
ELA	8	A	Claim 1 Literary	28	0.80
ELA	8	A	Claim 2	28	0.76
ELA	8	B	Claim 1	37	0.84
ELA	8	B	Claim 1 Information	18	0.72
ELA	8	B	Claim 1 Literary	19	0.73
ELA	8	B	Claim 2	20	0.68
ELA	8	C	Claim 1	43	0.84
ELA	8	C	Claim 1 Information	19	0.70
ELA	8	C	Claim 1 Literary	24	0.75
ELA	8	C	Claim 2	25	0.77
ELA	8	D	Claim 1	38	0.83
ELA	8	D	Claim 1 Information	21	0.73
ELA	8	D	Claim 1 Literary	17	0.69
ELA	8	D	Claim 2	17	0.68
ELA	8	E	Claim 1	39	0.81
ELA	8	E	Claim 1 Information	27	0.73
ELA	8	E	Claim 1 Literary	12	0.64
ELA	8	E	Claim 2	25	0.73
ELA	8	F	Claim 1	48	0.85
ELA	8	F	Claim 1 Information	21	0.71
ELA	8	F	Claim 1 Literary	27	0.77
ELA	8	F	Claim 2	28	0.75
ELA	8	G	Claim 1	44	0.84
ELA	8	G	Claim 1 Information	17	0.67
ELA	8	G	Claim 1 Literary	27	0.78
ELA	8	G	Claim 2	29	0.79
ELA	8	H	Claim 1	41	0.82
ELA	8	H	Claim 1 Information	18	0.69
ELA	8	H	Claim 1 Literary	23	0.71
ELA	8	H	Claim 2	29	0.76

Table E.14: Marginal Scaled Score Reliability for Grade 10 ELA Claim Scores

Subject	Grade	Form	Claim	Max Score	Reliability
ELA	10	A	Claim 1	44	0.86
ELA	10	A	Claim 1 Information	19	0.75
ELA	10	A	Claim 1 Literary	25	0.76
ELA	10	A	Claim 2	27	0.75
ELA	10	B	Claim 1	51	0.88
ELA	10	B	Claim 1 Information	29	0.82
ELA	10	B	Claim 1 Literary	22	0.74
ELA	10	B	Claim 2	23	0.70
ELA	10	C	Claim 1	51	0.85
ELA	10	C	Claim 1 Information	21	0.74
ELA	10	C	Claim 1 Literary	30	0.77
ELA	10	C	Claim 2	21	0.69
ELA	10	D	Claim 1	51	0.86
ELA	10	D	Claim 1 Information	20	0.72
ELA	10	D	Claim 1 Literary	31	0.78
ELA	10	D	Claim 2	34	0.78
ELA	10	E	Claim 1	42	0.84
ELA	10	E	Claim 1 Information	18	0.72
ELA	10	E	Claim 1 Literary	24	0.73
ELA	10	E	Claim 2	29	0.75
ELA	10	F	Claim 1	46	0.84
ELA	10	F	Claim 1 Information	27	0.78
ELA	10	F	Claim 1 Literary	19	0.65
ELA	10	F	Claim 2	36	0.77
ELA	10	G	Claim 1	48	0.86
ELA	10	G	Claim 1 Information	27	0.76
ELA	10	G	Claim 1 Literary	21	0.76
ELA	10	G	Claim 2	35	0.78
ELA	10	H	Claim 1	50	0.87
ELA	10	H	Claim 1 Information	21	0.75
ELA	10	H	Claim 1 Literary	29	0.79
ELA	10	H	Claim 2	33	0.75

Part VIII

Glossary, List of Figures and Tables, and References

Glossary of Assessment Terms

THE TABLE BELOW defines some of the assessment terms used in this technical manual. Although most of these terms are common within the assessment community, some are technical in nature.

Term	Definition
Ability	In IRT scaling, the construct measured by an exam—e.g., the KAP ELA test measures a student's ELA ability. A student who answers more items on that test correctly demonstrates greater ELA ability than a student who answers fewer items correctly.
Alternate Forms	Two or more versions of a test that are considered exchangeable. In other words, they measure the same constructs in the same ways, are intended for the same purposes, and are administered using the same directions.
Average	A general reference to the central tendency of a set of scores. It often refers to the arithmetic mean, which is determined by adding all the scores in a set and then dividing the result by the total number of scores (the n-count). Sometimes, *average* refers to the median, which is the middle score in a distribution.
Bias	Any source of systematic error in a test score. In discussing test fairness, bias refers to construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers (e.g., gender groups).
Claim Score	A set of test items that measure the same contextual area. On KAP score reports, claim scores reflect the examinee's performance for these sets of related items.
Constructed-Response Item (CR)	An item format where examinees create their own response. This may be done in a variety of ways, as long as it is appropriate for the construct being measured (e.g., a written paragraph, a created table or graph, a formulated calculation, or simply providing a short answer). CR items differ from item types that require students to select an answer choice from a supplied set of options, such as in a multiple-choice (MC) item.
Content Validity Evidence	Evidence that a test provides an appropriate sampling of a content domain (e.g., a state's Grade 6 mathematics curriculum). Content validity evidence would demonstrate that the assessment's components sample the domain's knowledge, skills, objectives, and processes in appropriate ways and proportions.
Criterion-Referenced Interpretation	An interpretation of a test score with respect to an expected level of mastery, educational objective, or standard. This type of interpretation provides information about what a student knows or can do with respect to a given content area.

Continued on next page

Term	Definition
Cut Score	A specified point on a score scale where scores at or above that point are interpreted or acted upon differently from scores below that point (e.g., a score designated as the minimum level of performance needed to pass a competency test). More than one cut score can be set for a test. The KAP uses three cut scores to place students into one of four performance levels.
Decision Consistency	The extent to which classifications based on one test's scores match the decisions based on scores from a second, parallel form of the same test. Decision consistency is often expressed as the proportion of examinees who are classified the same way by the two tests.
Differential Item Functioning (DIF)	An item exhibits DIF when students with the same ability level (usually based on their total test score) but different group memberships (e.g., male vs. female) do not have the same probability of answering the item correctly.
Distractor	An incorrect answer option to a multiple-choice item (also called a foil).
Equating	The strongest of several linking methods used to establish comparability between scores from multiple tests. Equated test scores can be considered exchangeable. The criteria needed to refer to a linkage as equating are strong and technically complex (equal construct and precision, equity, and invariance). In simple terms, it should not matter to a student which form of the equated tests he or she takes.
Error of Measurement	The amount by which an observed score differs from the true score, a hypothetical score that would be received if there were no errors of measurement.
Field-Test Item (FT)	A newly developed item that ready to be tried out to determine its statistical properties. Each KAP test form includes a set of FT items, but these FT items do not contribute to student scores.
Key	The correct-response option, or answer, to a test item.
Linking	A generic term that refers to one of a number of processes by which scores from one or more tests are made comparable to each other to some degree. Linking subsumes several terms, including equating, calibration, scale alignment, and prediction. Equating is associated with the strongest degree of comparability (exchangeable scores). Other score linkages may be strong but fail to meet one or more of the strict criteria required of equating.
Mean	A measure of central tendency in a score distribution found by adding all the score values in a distribution and dividing that sum by the total number of scores. The mean of this number set $\{21, 22, 23, 24, 30\}$ is 24. The mean can be influenced by scores that have extreme values.
Median	A measure of central tendency in a score distribution. It is the middle score in a set of rank-ordered observations. The median divides the distribution into two equal parts such that each part contains 50 percent of the observations in a data set. More simply put, half of the scores are below the median value and half of the scores are above the median value. The median for this ranked set of scores $\{21, 22, 23, 24, 30\}$ is 23.

Continued on next page

Term	Definition
N-count	The number of observations in a distribution of student scores or in a particular group. The n-count is sometimes simply designated as n. Examples of n include the number of students tested, the number of students tested from a specific subpopulation (e.g., females), or the number of students who attained a specific score. In this number set $\{23, 32, 56, 65, 78, 87\}$, $n = 6$.
Operational Item	An item in a test form that counts toward student total test score. The KAP tests use multiple forms for each grade-level subject-area test, and each form is composed of operational and (non-operational) field-test items.
P-value	An index that indicates an item's difficulty for some specified group (such as students at a particular grade level). The p-value is calculated as the proportion of students in the group who answer an item correctly, and it ranges from 0.0 to 1.0. Lower values correspond to more difficult items and higher values correspond to easier items. A p-value is usually provided for a multiple-choice item or an item of another format worth one point. For items worth more than one point, difficulty on a p-value-like scale can be estimated by dividing the mean item score by the maximum number of points possible for the item.
Performance Level Descriptor	Descriptions of levels of competency in a particular content area, usually represented as sequential categories that are labeled from lower levels to upper levels. Performance level descriptors constitute broad ranges for classifying performance. The exact labeling of these categories and their narrative descriptions usually vary from one assessment or testing program to another.
Point-Biserial Correlation	An item discrimination index in classical test theory. Point-biserial correlation is the correlation between a dichotomously scored item and a continuous criterion, usually represented by the total test score (or the corrected total test score with the reference item removed). It reflects the extent to which an item differentiates between high-scoring and low-scoring examinees. This discrimination index ranges from -1.00 to +1.00. The closer the index is to +1.00, the better the item's ability to separate low scoring students from high scoring students.
Reliability	The degree to which test scores for a group of examinees are consistent over exchangeable replications of an assessment procedure, and therefore, considered dependable or repeatable for an individual examinee. A test that produces highly consistent, stable results (i.e., one that is relatively free from random error) is said to be highly reliable. The reliability of a test is typically expressed as a reliability coefficient or by the standard error of measurement derived by that coefficient.
Reliability Coefficient	A statistical index that reflects the degree to which scores are free from random measurement error. Theoretically, the reliability coefficient expresses the consistency of test scores as the ratio of true score variance to total score variance (true score variance plus error variance). The closer the index is to its maximum value (1.0), the greater the reliability of the test.

Continued on next page

Term	Definition
Scaled Score	A mathematical transformation of a raw score which is developed through a process called scaling. Several different methods of scaling exist, but each is intended to provide a continuous and meaningful scale across different forms of a test.
Selected-Response Item	An item format that requires the test taker to select a response from a group of possible choices, one of which is the correct answer to the question.
Standard Deviation (SD)	A statistic that indicates the spread, or dispersion, of a set of scores. The value of a standard deviation is always greater than or equal to zero. If all of the scores in a set are identical, the standard deviation equals zero. The greater the difference in the score values, the greater the standard deviation.
Standard Error of Measurement (SEM)	The amount observed scores are expected to fluctuate around a true score (a hypothetical score that would be received if there were no errors in the measurement). Across replications of a measurement procedure, a student's true score will not differ by more than plus or minus one standard error from the observed score about 68 percent of the time (assuming normally distributed errors). The SEM is frequently used to determine the precision of a student's score in actual score units by establishing a confidence band around a student's score.
Standard Setting	A procedural event used to determine an assessment's cut scores, which are in turn used to determine if a students' test performance reflects a given performance standard. Standard setting methods vary, but most methods use a panel of educators to operationalize the level of achievement students must demonstrate in order to be categorized within a particular performance level.
Validity	The degree to which accumulated evidence supports the test score interpretations based on the intended uses of the test.

List of Figures

List of Tables

References

Allman, Carol B. 2004. "Test Access: Making Tests Accessible for Students with Visual Impairments: A Guide for Test Publishers, Test Developers, and State Assessment Personnel." *Louisville, Kentucky, June*.

Angoff, William H. 1984. *Scales, Norms, and Equivalent Scores*. Educational testing service.

Association, American Educational Research, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, and Psychological Testing (US). 1999. *Standards for Educational and Psychological Testing*. Amer Educational Research Assn.

———. 2013. *Standards for Educational and Psychological Testing*. Amer Educational Research Assn.

Birnbaum, Allan. 1968. "Some Latent Train Models and Their Use in Inferring an Examinee's Ability." *Statistical Theories of Mental Test Scores*. Addison-Wesley, 395–479.

Brennan, RL. 2004. "BB-CLASS: A Computer Program That Uses the Beta-Binomial Model for Classification Consistency and Accuracy (Version 1.0)(CASMA Research Report No. 9)." *Computer Software and Manual]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.(www. Education. Uiowa. Edu/casma)*.

Brennan, Robert L. 1989. *Methodology Used in Scaling: The ACT Assesment and P-ACT+*. ACT.

———. 1998. "Misconceptions at the Intersection of Measurement Theory and Practice." *Educational Measurement: Issues and Practice* 17 (1). Wiley Online Library: 5–9.

Buja, Andreas, and Nermin Eyuboglu. 1992. "Remarks on Parallel Analysis." *Multivariate Behavioral Research* 27 (4). Taylor & Francis: 509–40.

Cai, Li, and RJ Wirth. 2013. "FlexMIRT: Flexible Multilevel Multidimensional Item Analysis and Test Scoring."

Chen, Wen-Hung, and David Thissen. 1997. "Local Dependence Indexes for Item Pairs Using Item Response Theory." *Journal of Educational and Behavioral Statistics* 22 (3). Sage Publications: 265–89.

Cook, Linda L, and Daniel R Eignor. 1991. "IRT Equating Methods." *Educational Measurement: Issues and Practice* 10 (3). Wiley Online Library: 37–45.

Cronbach, Lee J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16 (3). Springer: 297–334.

Cronbach, Lee J, and Richard J Shavelson. 2004. "My Current Thoughts on Coefficient Alpha and Successor Procedures." *Educational*

and Psychological Measurement 64 (3). Sage Publications: 391–418.

Cronbach, Lee J, and RL Thorndike. 1971. "Educational Measurement." *Test Validation*. Wiley Online Library, 443–507.

Davier, Alina A von, and Christine Wilson. 2008. "Investigating the Population Sensitivity Assumption of Item Response Theory True-Score Equating Across Two Subgroups of Examinees and Two Test Formats." *Applied Psychological Measurement* 32 (1). Sage Publications: 11–26.

Dorans, Neil J, and Paul W Holland. 2000. "Population Invariance and the Equatability of Tests: Basic Theory and the Linear Case." *Journal of Educational Measurement* 37 (4). Wiley Online Library: 281–306.

Dorans, Neil J, Alicia P Schmitt, and Carole A Bleistein. 1992. "The Standardization Approach to Assessing Comprehensive Differential Item Functioning." *Journal of Educational Measurement* 29 (4). Wiley Online Library: 309–19.

Dorans, NJ, PW Holland, DT Thayer, and K Tateneni. 2003. "Invariance of Score Linking Across Gender Groups for Three Advanced Placement Program Examinations." *Population Invariance of Score Linking: Theory and Applications to Advanced Placement Program Examinations*, 79–118.

d'Agostino, Ralph B. 1998. "Tutorial in Biostatistics: Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group." *Stat Med* 17 (19): 2265–81.

Feldt, LS, and RL Brennan. 1989. "Reliability in Educational Measurement." New York, NY: American Council on Education.

Frisbie, David A. 2005. "Measurement 101: Some Fundamentals Revisited." *Educational Measurement: Issues and Practice* 24 (3). Wiley Online Library: 21–28.

Gulliksen, Harold. 1950. *Theory of Mental Tests*. John Wiley; Sons.

Haertel, Edward H. 2006. "Reliability." *Educational Measurement*. Praeger Pub Text, 65–110.

Hambleton, Ronald K, and Melvin R Novick. 1973. "Toward an Integration of Theory and Method for Criterion-Referenced Tests." *Journal of Educational Measurement*. JSTOR, 159–70.

Hambleton, Ronald K, and H Jane Rogers. 1986. "Evaluation of the Plot Method for Identifying Potentially Biased Test Items." Kluwer Academic Publishers.

Hambleton, Ronald K, H Swaminathan, and H Jane Rogers. 1991. "Fundamentals of Item Response Theory (Measurement Methods for the Social Sciences Series, Vol. 2)." London: SAGE.

Hanson, Bradley A, and Robert L Brennan. 1990. "An Investigation of Classification Consistency Indexes Estimated Under Alternative Strong True Score Models." *Journal of Educational Measurement*

27 (4). Wiley Online Library: 345–59.

Harvill, Leo M. 1991. "Standard Error of Measurement." *Educational Measurement: Issues and Practice* 10 (2). Wiley Online Library: 33–41.

Hoover, HD. 1984. "The Most Appropriate Scores for Measuring Educational Development in the Elementary Schools: GE's." *Educational Measurement: Issues and Practice* 3 (4). Wiley Online Library: 8–14.

Horn, John L. 1965. "A Rationale and Test for the Number of Factors in Factor Analysis." *Psychometrika* 30 (2). Springer: 179–85.

Houts, CR, and L Cai. 2013. "FlexMIRT User's Manual Version 2.0: Flexible Multilevel Multidimensional Item Analysis and Test Scoring." *Chapel Hill, NC: Vector Psychometric Group. OpenURL*.

Huynh, Huynh. 1976. "On the Reliability of Decisions in Domain-Referenced Testing." *Journal of Educational Measurement* 13 (4). Wiley Online Library: 253–64.

Jodoin, Michael G, and Mark J Gierl. 2001. "Evaluating Type I Error and Power Rates Using an Effect Size Measure with the Logistic Regression Procedure for DIF Detection." *Applied Measurement in Education* 14 (4). Taylor & Francis: 329–49.

Kaiser, Henry F. 1960. "The Application of Electronic Computers to Factor Analysis." *Educational and Psychological Measurement*. Sage Publications.

Kansas Assessment Examiner's Manual 2015-2016. 2015. 900 SW Jackson Topeka, Kansas 66612-1212: Kansas State Department of Education.

Karkee, Thakur, Dong-In Kim, and Kevin Fatica. 2010. "Comparability Study of Online and Paper and Pencil Tests Using Modified Internally and Externally Matched Criteria." In *Annual Meeting of the American Educational Research Association*.

Lane, Suzanne. 1999. "Validity Evidence for Assessments." Edward F. Reidy Interactive Lecture Series, The National Center for the Improvement of Educational Assessment, Providence, RI.

Lane, Suzanne, and Clement A Stone. 2002. "Strategies for Examining the Consequences of Assessment and Accountability Programs." *Educational Measurement: Issues and Practice* 21 (1). Wiley Online Library: 23–30.

Lewis, Daniel M, Harold C Mitzel, and Donald R Green. 1996. "Standard Setting: A Bookmark Approach." In *DR Green (Chair), IRT-Based Standard Setting Procedures Utilizing Behavioral Anchoring. Symposium Conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment. Phoenix, AZ*.

Livingston, Samuel A, and Charles Lewis. 1995. "Estimating the Consistency and Accuracy of Classifications Based on Test Scores."

Journal of Educational Measurement 32 (2). Wiley Online Library: 179–97.

Lord, Frederic M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Routledge.

Mantel, Nathan, and William Haenszel. 1959. "Statistical Aspects of the Analysis of Data from Retrospective Studies." *J Natl Cancer Inst* 22 (4): 719–48.

McDonald, Roderick P. 1979. "The Structural Analysis of Multivariate Data: A Sketch of a General Theory." *Multivariate Behavioral Research* 14 (1). Taylor & Francis: 21–38.

Messick, Samuel. 1989. "Educational Measurement." In, edited by Robert L. Linn. Praeger.

Moses, Tim, Weiling Deng, and Yu-Li Zhang. 2010. "The Use of Two Anchors in Nonequivalent Groups with Anchor Test (NEAT) Equating." *ETS Research Report Series* 2010 (2). Wiley Online Library: i–i33.

Orlando, Maria, Cathy D Sherbourne, and David Thissen. 2000. "Summed-Score Linking Using Item Response Theory: Application to Depression Measurement." *Psychological Assessment* 12 (3). American Psychological Association: 354.

Petersen, Nancy S, Michael J Kolen, and H Dv Hoover. 1989. "Educational Measurement." In, edited by Robert L. Linn. Praeger.

Qualls, Audrey L. 1995. "Estimating the Reliability of a Test Containing Multiple Item Formats." *Applied Measurement in Education* 8 (2). Taylor & Francis: 111–20.

Reckase, Mark D. 1979. "Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications." *Journal of Educational and Behavioral Statistics* 4 (3). Sage Publications: 207–30.

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1). Biometrika Trust: 41–55.

Rosenbaum, Paul R. 2002. *Observational Studies*. Springer.

Rosseel, Yves. 2012. "Lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software* 48 (2): 1–36.

Rubin, Donald B. 2006. *Matched Sampling for Causal Effects*. Cambridge University Press.

Samejima, Fumiko. 1969. "Estimation of Latent Ability Using a Response Pattern of Graded Scores." *Psychometrika Monograph Supplement*.

———. 1997. "Graded Response Model." In *Handbook of Modern Item Response Theory*, 85–100. Springer.

Spearman, Charles. 1904. "The Proof and Measurement of Association Between Two Things." *The American Journal of Psychology* 15 (1). JSTOR: 72–101.

———. 1910. "Correlation Calculated from Faulty Data." *British Journal of Psychology, 1904-1920* 3 (3). Wiley Online Library: 271–95.

Swaminathan, Hariharan, and H Jane Rogers. 1990. "Detecting Differential Item Functioning Using Logistic Regression Procedures." *Journal of Educational Measurement*. JSTOR, 361–70.

Thompson, S, CJ Johnston, and Martha L Thurlow. 2002. "Universal Design Applied to Large Scale Assessments." National Center on Educational Outcomes Synthesis Report.

Traub, Ross E. 1994. *Reliability for the Social Sciences: Theory and Applications*. Vol. 3. Sage.

Way, Walter D, Chow-Hong Lin, and Jadie Kong. 2008. "Maintaining Score Equivalence as Tests Transition Online: Issues, Approaches and Trends." In *Annual Meeting of the National Council on Measurement in Education, New York, NY*.

Yen, Wendy M. 1993. "Scaling Performance Assessments: Strategies for Managing Local Item Dependence." *Journal of Educational Measurement* 30 (3). Wiley Online Library: 187–213.