Kansas Assessment Program
Technical Manual
2017

University of Kansas Achievement & Assessment Institute

November 2017

Table of Contents

I. Statewide System of Standards and Assessments 1
I.1. State Adoption of Academic Content Standards for All Students 2
I.2. Coherent and Rigorous Academic Content Standards 2
I.2.1. Process and timeline. 2
I.2.2. Convergence and divergence with national standards. 3
I.2.3. Standards review committees. 4
I.3. Required Assessments 5
I.4. Policies for Including All Students in Assessments 5
I.5. Participation Data 6
II. Assessment System Operations 7
II.1. Assessment Framework of the Assessed Grades 7
II.2. Test Design and Development 12
II.2.1. Test blueprints. 14
II.2.2. Test design. 14
II.2.3. Operational test construction. 16
II.2.3.1. ELA and mathematics test construction guidelines. 17
II.2.3.2. Science test construction guidelines. 17
II.2.4. Item pool evaluation. 17
II.2.4.1. Alignment study of adaptive test item pool. 17
II.2.4.2. Item count by content standard. 18
II.2.4.3. Item statistics. 19
II.3. Item Development 20
II.3.1. Passage selection and review. 20
II.3.2. Item writers. 22
II.3.3. Item-writing training. 22
II.3.3.1. General guidelines. 22
II.3.3.2. Content guidelines. 23
II.3.3.3. Format guidelines. 23
II.3.3.4. Structure guidelines. 23
II.3.3.5. Stem construction guidelines. 23
II.3.3.6. Answer-choice development guidelines 23

II.3.3.7. Accessibility guidelines..... 24
II.3.3.8. Bias and sensitivity guidelines. 24
II.3.4. Item writing. 24
II.3.5. Item reviewers. 25
II.3.6. Item review. 26
II.3.7. Universal design (UD) in test development. 27
II.3.8. Field testing. 27
II.3.9. Field test data analysis..... 27
II.3.10. Data review. 27
II.4. Test Administration..... 28
II.4.1. Test administration and security training. 28
II.4.2. Monitoring test administration. 28
II.5. Systems for Protecting Data Integrity and Privacy 29
III. Technical Quality—Validity..... 31
III.1. Overall Validity, Including Validity Based on Content..... 31
III.1.1. Content validity. 31
III.2. Validity Based on Cognitive Process..... 33
III.3. Validity Based on Internal Structure..... 34
III.3.1. Internal construct. 34
III.3.2. IRT and model assumptions..... 34
III.3.2.1. Samples. 35
III.3.2.2. Missing data. 35
III.3.2.3. Excluded items. 36
III.3.2.4. IRT models..... 36
III.3.2.5. Evaluating IRT assumptions. 36
III.3.2.5.1. IRT model fit. 36
III.3.2.5.2. Unidimensionality. 37
III.3.2.5.3. Local independence. 37
III.3.2.5.4. Invariance. 37
III.3.3. Differential item functioning (DIF). 39
III.4. Validity Based on Relationships to Other Variables 40
III.4.1. Relationships among KAP subjects. 41

III.4.2. Relationships between scale scores and demographic variables. 41
III.4.3. Relationships between KAP scores and ACT scores. 42
III.4.4. Relationships between the KAP assessment and the National Assessment of Educational Progress (NAEP). 44
III.5. Survey Results for Validity Evidence 46
IV. Technical Quality—Others 48
IV.1. Reliability 48
IV.1.1. Test reliability. 48
IV.1.2. Classification consistency and accuracy. 49
IV.1.3. Subgroup reliability. 51
IV.1.4. Path reliability. 51
IV.1.2. Subscore reliability. 52
IV.2. Fairness and Accessibility 54
IV.3. Full Performance Continuum 55
IV.3.1. Classical item statistics. 55
IV.3.2. IRT item statistics. 58
IV.3.3. Cognitive complexity. 60
IV.4. Scoring and Scaling 61
IV.4.1. Item scoring. 61
IV.4.2. Test scoring. 61
IV.4.3. Scaling. 61
IV.4.3.1. Scale transformation and cut scores. 62
IV.4.3.2. ELA and mathematics scale transformation. 62
IV.4.3.3. Properties of scaled scores. 64
IV.4.4. Operational test results. 64
IV.5. Multiple Assessment Forms 70
IV.5.1. Within-year linking design. 70
IV.5.2. Cross-year linking design. 70
IV.5.3. Linking procedure. 71
IV.6. Technical Analysis and Ongoing Maintenance 71
IV.6.1. Model change. 71
IV.6.2. Elimination of hand-scored tasks. 72
IV.6.3. Elimination of Listening. 73

V. Inclusion of All Students 75
V.1. Procedures for Including Students with Disabilities 75
V.2. Accommodations 76
V.3. Frequency of Accommodation Use 76
VI. Academic Achievement Standards and Reporting 78
VI.1. State Adoption of Academic Achievement Standards for All Students 78
VI.2. Achievement Standard Setting 78
VI.2.1. Overview of the Bookmark method. 78
VI.2.2. The ordered item booklet (OIB). 79
VI.2.3. Panelist recruiting process. 79
VI.2.4. Performance level descriptors (PLDs). 80
VI.2.5. Standard-setting procedure. 80
VI.2.5.1. Training session. 80
VI.2.5.2. Completing the science exam. 81
VI.2.5.3. “Just-barely” student activity and discussion. 82
VI.2.5.4. Bookmark practice. 82
VI.2.5.5. Identify operational item knowledge and skills. 83
VI.2.5.6. Setting cut score: Round 1. 83
VI.2.5.7. Setting cut score: Round 2. 84
VI.2.5.8. Setting cut score: Round 3. 84
VI.2.5.9. Articulation training and articulation 86
VI.3. Challenging and Aligned Academic Achievement Standards 87
VI.4. Reporting 87
VI.4.1. Group masking 88
VI.4.2. Student reports. 88
Timeline for delivering student reports 89
VI.4.3. School and district reports. 89
VI.4.4. Interpretive guides. 89
VI.4.5. Letters from the Commissioner of Education. 89
VII. References 90
VIII. Appendix A: Test Administration and Security Training 93
IX. Appendix B: Conditional Standard Error of Measurement (CSEM) 96

X. Appendix C: Frequency Distribution and CSEM of Scale Scores 102
XI. Appendix D: Subgroup Reliability and Performance 141
XII. Appendix E: Path Reliability 150
XIII. Appendix F: Scale Score Frequency Distribution 150
XIV. Appendix G: 2017 Kansas Assessment Survey Results 156
XV. Appendix H: Science Performance Level Descriptors (PLDs) 163
XVI. Appendix I: Science Standard-Setting Meeting Agenda 166
Tuesday, June 20, 2017 166
XVII. Appendix J: Participant Survey 167
XVIII. Appendix K: Confidentiality Agreement and Statement of Original Work 168
XIX. Appendix L: Readiness Form 169
XX. Appendix M: Evaluation Form 170
XXI. Appendix N: Articulation Session Evaluation Form 180
XXII. Appendix O: Score Reports 181
XXIII. Appendix P: Letters from the Commissioner of Education 191

Table of Tables

Table I-1. Number and Percentage of Enrolled Students Tested by Subject Test and Grade 6
Table II-1. ELA Claims and Targets 7
Table II-2. Mathematics Claims and Targets 8
Table II-3. Science Grade 5 Claims and Targets 8
Table II-4. Science Grade 8 Claims and Targets 9
Table II-5. Science Grade 11 Claims and Targets 10
Table II-6. HGSS Standards and Benchmarks 11
Table II-7. Development Timeline for the KAP Assessment 13
Table II-8. Test Blueprint by Subject and Claim/Category 14
Table II-9. Test Design for the KAP Assessment 15
Table II-10. Number and Difficulty of Blocks for the KAP Assessment 16
Table II-11. Percentages of Items by Claims 19
Table II-12. Pathways of Multistage Design for ELA and Mathematics 19
Table III-1. Sample Size for Concurrent Calibration by Grade for Science 35
Table III-2. Science Misfit Results by Grade 37
Table III-3. Science Item-Parameter Correlations between Female and Male Samples 39
Table III-4. ELA DIF Item Count by Grade 40
Table III-5. Mathematics DIF Item Count by Grade 40
Table III-6. Science DIF Item Count by Grade 40
Table III-7. Correlations Among ELA, Mathematics, and Science Scores 41
Table III-8. Correlations Between Scale Scores and Demographic Groups for ELA 42
Table III-9. Correlations Between Scale Scores and Demographic Groups for Mathematics 42
Table III-10. Correlations Between Scale Scores and Demographic Groups for Science 42
Table III-11. Correlations Among KAP and ACT Scores (N = 5,369) 43
Table IV-1. Test Reliability by Grade and Subject 49
Table IV-2. Cross-Tabulation of Classification Consistency 50
Table IV-3. Cross-Tabulation of Classification Accuracy 50
Table IV-4. Classification Consistency and Accuracy by Subject and Grade 51
Table IV-5. ELA Grade 3 Path Reliability 52
Table IV-6. Additional Subscores for ELA by Grade 53
Table IV-7. Summary of Subscore Reliability and Classification Consistency and Accuracy by Subject 54
Table IV-8. Summary Statistics for Classical Item Difficulties for ELA 56
Table IV-9. Summary Statistics for Classical Item Difficulties for Mathematics 56
Table IV-10. Summary Statistics for Classical Item Difficulties for Science 56
Table IV-11. Summary Statistics for Classical Item Discrimination for ELA 57
Table IV-12. Summary Statistics for Classical Item Discrimination for Mathematics 58
Table IV-13. Summary Statistics for Classical Item Discrimination for Science 58
Table IV-14. Summary Statistics for IRT Item Difficulty for ELA 59
Table IV-15. Summary Statistics for IRT Item Difficulty for Mathematics 59
Table IV-16. Summary Statistics for IRT Item Difficulty for Science 59

Table IV-17. Summary Statistics for IRT Item Discrimination for ELA 59
Table IV-18. Summary Statistics for IRT Item Discrimination for Mathematics 60
Table IV-19. Summary Statistics for IRT Item Discrimination for Science 60
Table IV-20. Number of Items by DOK Level, Subject, and Grade 61
Table IV-21. ELA Cut Scores 63
Table IV-22. Mathematics Cut Scores 63
Table IV-23. Science Cut Scores 63
Table IV-24. ELA, Mathematics, and Science Scaling Constants 64
Table IV-25. Scaled-Score Descriptive Statistics by Grade for ELA 65
Table IV-26. Scaled-Score Descriptive Statistics by Grade for Mathematics 65
Table IV-27. Scaled-Score Descriptive Statistics by Grade for Science 65
Table IV-28. Percentage of Students in Each Performance Level by Subject and Grade 66
Table IV-29. Longitudinal Scaled-Score Trend for ELA 68
Table IV-30. Longitudinal Scaled-Score Trend for Mathematics 68
Table IV-31. Mathematics Scale Scores by Grade: A Comparison Between Tests with and Without Performance Tasks 72
Table IV-32. Reliability Difference Between Mathematics Tests with and Without Performance Tasks 73
Table IV-33. ELA Scale Scores by Grade: A Comparison Between Tests with and Without Listening Items 73
Table IV-34. Reliability Difference Between ELA Tests with and Without Performance Tasks 74
Table V-1. Available Nonreported and Reported KAP Accommodations 76
Table V-2. Frequency of Accommodation Requests by Grade 77
Table VI-1. Summary of Science Panelists' Demographic Information 81
Table VI-2. Rounds 1–3 Medians of Bookmark Placements by Grade 85
Table VI-3. Summary of Panelists' Perceptions about Cut-Score Results in Evaluation Survey 86
Table VI-4. Articulation Data by Grade and Level 87

Table of Figures

Figure II-1. An example of Stage 2 block information curves. 20
Figure III-1. Science item-discrimination parameter scatter plot by grade. 38
Figure III-2. Science item-difficulty parameter scatter plot by grade. 39
Figure III-3. Grade 10 ELA trends across years: KAP vs. ACT. 43
Figure III-4. Grade 10 mathematics trends across years: KAP vs. ACT. 44
Figure III-5. Grade 4 ELA trend across years: KAP vs. NAEP. 45
Figure III-6. Grade 8 ELA trend across years: KAP vs. NAEP. 45
Figure III-7. Grade 4 mathematics trend across years: KAP vs. NAEP. 46
Figure III-8. Grade 8 mathematics trend across years: KAP vs. NAEP. 46
Figure IV-1. Performance-level results for ELA. 66
Figure IV-2. Performance-level results for mathematics. 67
Figure IV-3. Performance-level results for science. 67
Figure IV-4. Performance-level trend for ELA. G = grade. 68
Figure IV-5. Career-readiness trend for ELA. G = grade. 69
Figure IV-6. Performance-level trend for mathematics. G = grade. 69
Figure IV-7. Career-readiness trend for mathematics. G = grade 70

I. Statewide System of Standards and Assessments

The Kansas Assessment Program (KAP), a program of the Kansas State Board of Education (hereafter the State Board), is mandated by the Kansas State Legislature. In addition, the English language arts (ELA), mathematics, and science components of KAP also are used to comply with federal elementary and secondary education legislation. The four main purposes of KAP, as stated in the *Kansas Assessment Examiner's Manual 2016–2017* (hereafter the *Examiner's Manual*; Kansas State Department of Education [KSDE], 2017), are to accomplish the following:

- measure specific claims related to the Kansas College and Career Ready Standards (KCCRS);
- provide information for calculating Annual Measurable Objectives 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's or district's programs in ELA, mathematics, and science.

The state statutory authority behind KAP is Kan. Stat. Ann. §72-6479: School performance accreditation system; curriculum standards; student assessments; school site councils (2015). According to this statute, the State Board is mandated, in part, to complete the following:

- 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 for statewide assessments in the core academic areas of mathematics, science, reading, writing and social studies and determine performance levels on the statewide assessments.

KAP offers two summative assessments: the test for the general student population and the alternate assessment for students with significant cognitive disabilities. Additionally, KAP provides a language proficiency test for English learners (EL). This technical manual addresses the test for the general student population; The general population test will further be referred to as the KAP assessment. For the convenience of stakeholders, this manual follows the reporting structure recommended in the 2015 Assessment Peer Review Guidance (U.S. Department of Education, 2015).

I.1. State Adoption of Academic Content Standards for All Students

The state legislature mandates that KSDE review the Kansas curriculum standards every seven years. The State Board adopted the KCCRS for ELA and mathematics in 2010; for science in June 2013; and for history, government, and social studies (HGSS) in April 2013. The first operational administration of the KCCRS-aligned KAP assessments for ELA and mathematics was in 2015, HGSS in 2016, and science in 2017.

I.2. Coherent and Rigorous Academic Content Standards

Committees of Kansas educators and stakeholders developed and reviewed the standards in Kansas. These standards help schools prepare students by outlining knowledge and skills needed to pursue higher education or better careers and to compete in an increasingly competitive and global work environment. The KCCRS are Kansas's coherent and rigorous academic content standards, which adhere to the State Board's mission.

> The mission of Kansas State Board of Education is to prepare Kansas students for lifelong success through rigorous, quality academic instruction, career training and character development according to each student's gifts and talents. The Kansas CAN Vision is to Lead the World in the Success of Each Student (refer to http://www.ksde.org/Board).

I.2.1. Process and timeline. Under the direction of and feedback from Kansas educators, the KCCRS in ELA and mathematics 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 or mathematics KCCRS committee provided feedback to the CCSSO and other groups involved in the development process; this feedback was incorporated into subsequent drafts of the CCSS. In September 2010, the standards for ELA and mathematics were presented to the State Board, which on October 10, 2010, adopted the KCCRS for ELA and mathematics for use in Kansas.

Kansas led the development of the Next Generation Science Standards (NGSS). Beginning in 2011, participating states and standards writers were recruited to start the development process. Between 2011 and 2013, writing teams and stakeholders reviewed and revised a series of drafts of the science standards, which included periods of public review. When the new standards were completed, the Kansas standards development committee thoroughly reviewed the document to verify that feedback from the Kansas review team was acknowledged and that the standards represented the best interests of Kansas students. In May 2013, the Kansas NGSS review committee recommended the KCCRS for science to the State Board, which adopted the standards on June 11, 2013, after a month of deliberation.

The development of Kansas HGSS standards was undertaken by a committee of Kansas educators and stakeholders in May 2011 and culminated with adoption of the standards by the State Board on April 16, 2013. From the outset, the goal of the HGSS standards committee was to create a document that would emphasize and encourage the application of content in authentic

situations, rather than a traditional approach to HGSS standards that focuses on dates and minutiae. To this end, the final standards represent methods of thinking rather than a document to be used as a scope and sequence. The Mission Statement in the HGSS content standards reads "[t]he Kansas Standards for History, Government, and Social Studies prepare students to be informed, thoughtful, engaged citizens as they enrich their communities, state, nation, world, and themselves." (KSDE, 2013, p. 5)

The drafting of the content standards was an iterative process, moving from the committee to public comment and review and finally back to the committee. In total, the document went through three cycles of public review and revision before it was submitted to the State Board in October 2012 for review and feedback. The HGSS committee incorporated the additional changes recommended by the State Board and presented the standards for adoption in March 2013. The State Board adopted the standards in April 2013.

I.2.2. Convergence and divergence with national standards. According to the CCSS Initiative, 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. (CCSS Initiatives, 2010, p. 10)

The key difference between the national CCSS and Kansas's KCCRS is the Kansas 15%, the purpose of which is to emphasize concepts and teaching philosophies that are important in Kansas. Although most of the Kansas concepts are mentioned in the CCSS, KSDE wanted to highlight the importance of each one by including the concepts and teaching philosophies in the KCCRS. As part of the Kansas 15%, KSDE added the anchor standards for literacy learning, as well as four other anchor standards—two in reading and two in writing (KSDE, 2010a).

For mathematics, the Kansas additions to the CCSS were for probability and statistics and also algebraic patterning. These two topics were left for each school and/or district to decide how to incorporate them (KSDE, 2010b).

The development of the NGSS was led by Kansas; thus, the KCCRS for science closely align to the NGSS. The NGSS are based on the *Framework for K–12 Science Education* developed in 2012 by the National Research Council of the National Academies. However, the intent of the NGSS is to put the *Framework* into practice by coupling the practice with content, providing performance expectations while leaving curricular and instructional decisions to states and educators, and evaluating students on the degree of understanding of a full discipline core idea. The NGSS were developed because

> the world has changed dramatically in the 15 years since state science education standards' guiding documents were developed. Since that time, many advances have occurred in the fields of science and science education, as well as in the innovation-driven economy. The U.S. has a leaky K–12 science, technology, engineering and mathematics (STEM) talent pipeline, with too few students entering STEM majors and careers at every level—from those with relevant postsecondary certificates to PhD's. We need new science standards that stimulate and build interest in STEM.
>
> The current education system can't successfully prepare students for college, careers and citizenship unless we set the right expectations and goals. While standards alone are no silver bullet, they do provide the necessary foundation for local decisions about curriculum, assessments, and instruction.
>
> Implementing the NGSS will better prepare high school graduates for the rigors of college and careers. In turn, employers will be able to hire workers with strong science-based skills—not only in specific content areas, but also with skills such as critical thinking and inquiry-based problem solving. (Next Generation Science Standards, 2013, p. 1 of Introduction)

The mission of the KCCRS for HGSS, as described in the *Kansas Standards for History, Government, and Social Studies*, is to "prepare students to be informed, thoughtful, engaged citizens as they enrich their communities, state, nation, world, and themselves" (KSDE, 2013, p. 5). To develop the KCCRS for HGSS, the standard writing committee

> reviewed other state and national standards, researched best instructional practices, and gathered input from professionals and citizens in order to define what Kansas students should be able to know and to do in history, civics/government, geography, and economics. The committee responded to feedback on earlier versions throughout the current process. This revised document focuses on discipline-specific habits of mind that encourage the application of content in authentic situations rather than specific content, and is intended as a framework for curriculum, instruction, assessment, and teacher preparation. (KSDE, 2013, p. 6)

I.2.3. Standards review committees. Committee members involved in the development of the Kansas additions to the CCSS for ELA and mathematics were recruited from across the state. The ELA committee comprised 22 members, and the mathematics committee comprised nine members; most members were K–12 educators. Additionally, two representatives from postsecondary education were recruited for each subject.

The Kansas review team and the Kansas science education committee, a subcommittee of the review team, reviewed the KCCRS for science. The review team included 60 members from across the state and comprised K–12 science educators, postsecondary science professors, and business and industry professionals. The subcommittee focused on finding ways to "build and

leverage relationships between P-20[1] educators and business and industry to build state-wide capacity for science education in Kansas" (Kansas Next Generation Science Education, http://community.ksde.org/Default.aspx?tabid=5407).

The panel of HGSS committee members came from the result of nominations from State Board members and the Commissioner of Education, as well as internal nominations from KSDE content staff who were familiar with top Kansas educators and community leaders. Committee members included representation from the community at large and also several state and national organizations. Their expertise included HGSS teaching and curriculum, special education, and educators of EL. The final committee comprised approximately 30 individuals from across the state and was facilitated by Donald Gifford of KSDE.

I.3. Required Assessments

The KAP assessment tests students in the subject areas of ELA, mathematics, science, and HGSS. The subject areas and grades tested are as follows:

- ELA in grades 3–8 and 10;
- mathematics in grades 3–8 and 10;
- science in grades 5, 8, and 11; and
- HGSS in grades 6, 8, and 11 (tested in even-numbered years, e.g., 2016, 2018, etc.).

I.4. Policies for Including All Students in Assessments

Kansas is committed to include all students in the KAP assessment. Students enrolled in Kansas public schools must take one of three tests: the KAP assessment, the English language proficiency test, or the alternate assessment. The EL students who are recent arrivals to the United States are required to take the KAP mathematics and science tests, but their results count only toward participation. They are not required to take the ELA or HGSS tests but must take the English language proficiency test.

Qualifying students with significant cognitive disabilities, typically no more than 1% of Kansas students, take the Dynamic Learning Maps® Alternate Assessment for ELA, mathematics, and science and a separate HGSS Alternate Assessment. Other special-needs students with Individualized Education Programs, 504 plans, or Student Intervention Team plans take the KAP assessment but can use accommodations consistent with their personal needs profiles (PNPs). If an unapproved accommodation is used (e.g., reading aloud to a student on the KAP ELA test), the student is considered "not tested." A detailed accommodation summary can be found in chapter V. Inclusion of All Students of this technical manual.

Only a few exemptions are granted to students. The exemptions include the following:

- students serving long-term suspension;
- students who were truant for more than two consecutive weeks at the time of testing;

[1] P-20 refers to the integrated education system that extends from preschool through higher education.

- students who experienced catastrophic illnesses or accidents during testing;
- students who moved during testing; and
- students who were incarcerated during testing.

I.5. Participation Data

In 2017, the KAP operational test was administered in ELA, mathematics, and science. Table I-1 shows the number and percentage of enrolled students who took each test in each grade. The tested rates of all grades other than high school are 99% or above; the high school tested rate for mathematics and science is 97% and for ELA is 98%.

Table I-1. Number and Percentage of Enrolled Students Tested by Subject Test and Grade

	Grade						
Subject test	3	4	5	6	7	8	HS
ELA							
No. enrolled	38,513	38,625	37,692	37,038	37,076	36,929	36,288
No. tested	38,287	38,355	37,478	36,813	36,791	36,618	35,415
Percentage tested	99%	99%	99%	99%	99%	99%	98%
Mathematics							
No. enrolled	38,560	38,660	37,726	37,075	37,102	36,971	36,327
No. tested	38,371	38,446	37,558	36,853	36,817	36,660	35,316
Percentage tested	99%	99%	99%	99%	99%	99%	97%
Science							
No. enrolled			37,723			36,963	34,835
No. tested			37,567			36,709	34,009
Percentage tested			99%			99%	97%

Note. HS = high school.

II. Assessment System Operations

The development of any test requires making many critical decisions regarding, for example, the content and cognitive complexity, as well as the appropriate scope, sequence, and progression of that content for particular subject areas. Other decisions are related to the number of points for each test and the proportion of those points for any subscores. These decisions are not made in isolation but must be reasonable across all grade levels of the assessment. Together, these decisions represent the constructs that a test measures. Critical test construction–related documents yielded from these decisions include development timeline and test blueprint. These documents guide the test construction process and products.

II.1. Assessment Framework of the Assessed Grades

KAP KCCRS content standards—except for HGSS, which labels its content hierarchy by standards and benchmarks—are defined for the purposes of assessment and are reported in two levels: claims and targets. Not all claims have a target sublevel. An item is aligned to only one claim or target. However, in 2017, some targets were combined for the purposes of additional subscore reporting; therefore, in some instances, an item was included in multiple subscores.

Tables II-1 through II-6 show the KCCRS assessment framework for the four KAP subjects. ELA and mathematics have the same claims and targets across grades. ELA has two claims: reading and writing. Mathematics has four claims, and all of its targets are under Claim 1. Science has the same claims but different targets across grades. HGSS standards and benchmarks are identical across grades.

Table II-1. ELA Claims and Targets

Claim	Claim label	Target
1	Reading – Literary and Informational Texts	Central ideas (Targets 2, 9) Word meanings (Targets 3, 10) Making and supporting conclusions or inferences (Targets 1, 4, 8, 11) Analysis within or across texts (Targets 5, 12) Text structures or features (Targets 6, 13) Language use (Targets 7, 14)
2	Writing	Revising narrative, informational, and argumentative texts (Targets 1, 3, 6) Vocabulary and language use (Target 8) Editing (Target 9)

Table II-2. Mathematics Claims and Targets

Claim	Claim label	Target
1	Concepts and procedures	Operations and algebraic thinking Number and operations in base ten Numbers and operations with fractions Measurement and data Geometry The number system Expressions and equations Statistics and probability Algebra
2	Problem solving	
3	Communicating reasoning	
4	Modeling and data analysis	

Table II-3. Science Grade 5 Claims and Targets

Claim	Claim label	Target
1	Physical science	A. Structure and properties of matter B. Engineering design in physical science
2	Life science	A. Matter and energy in organisms and ecosystems B. Engineering design in life science
3	Earth and space science	A. Earth's systems B. Space systems C. Engineering design in earth and space science

Table II-4. Science Grade 8 Claims and Targets

Claim	Claim label	Target
1	Physical science	A. Structure and properties of matter B. Chemical reactions C. Forces and interactions D. Energy E. Waves and electromagnetic radiation F. Engineering design in physical science
2	Life science	A. Structure, function, and information processing B. Matter and energy in organisms and ecosystems C. Interdependent relationships in ecosystems D. Growth, development, and reproduction of organisms E. Natural selection and adaptations F. Engineering design in life science
3	Earth and space science	A. Space systems B. History of the earth C. Earth's systems D. Weather and climate E. Human impacts F. Engineering design in earth and space science

Table II-5. Science Grade 11 Claims and Targets

Claim	Claim label	Target
1	Physical science	A. Structure and properties of matter B. Chemical reactions C. Forces and interactions D. Energy E. Waves and electromagnetic radiation F. Engineering design in physical science
2	Life science	A. Structure and function B. Matter and energy in organisms and ecosystems C. Interdependent relationships in ecosystems D. Inheritance and variation of traits E. Natural selection and evolution F. Engineering design in life science
3	Earth and space science	A. Space systems B. History of the earth C. Earth's systems D. Weather and climate E. Human sustainability F. Engineering design in earth and space science

Table II-6. HGSS Standards and Benchmarks

Standard	Benchmark
Choices have consequences.	1.1 The student will recognize and evaluate significant choices made by individuals, communities, states, and nations that have impacted our lives and futures. 1.2 The student will analyze the context under which choices are made and draw conclusions about the motivations and goals of the decision-makers. 1.3 The student will investigate examples of causes and consequences of particular choices and connect those choices with contemporary issues. 1.4 The student will use his/her understanding of choices and consequences to construct a decision-making process and to justify a decision.
Individuals have rights and responsibilities.	2.1 The student will recognize and evaluate the rights and responsibilities of people living in societies. 2.2 The student will analyze the context under which significant rights and responsibilities are defined and demonstrated, their various interpretations, and draw conclusions about those interpretations. 2.3 The student will investigate specific rights and responsibilities of individuals and connect those rights and responsibilities with contemporary issues. 2.4 The student will use his/her understanding of rights and responsibilities to address contemporary issues.
Societies are shaped by beliefs, ideas, and diversity.	3.1 The student will recognize and evaluate significant beliefs, contributions, and ideas of the many diverse peoples and groups and their impact on individuals, communities, states, and nations. 3.2 The student will draw conclusions about significant beliefs, contributions, and ideas, analyzing the origins and context under which these competing ideals were reached and the multiple perspectives from which they come. 3.3 The student will investigate specific beliefs, contributions, ideas, and/or diverse populations and connect those beliefs, contributions, ideas and/or diversity to contemporary issues. 3.4 The student will use his/her understanding of those beliefs, contributions, ideas, and diversity to justify or define how community, state, national, and international ideals shape contemporary society.
Societies experience continuity and change over time.	4.1 The student will recognize and evaluate continuity and change over time and its impact on individuals, institutions, communities, states, and nations.

Standard	Benchmark
	4.2 The student will analyze the context of continuity and change and the vehicles of reform, drawing conclusions about past change and potential future change. 4.3 The student will investigate an example of continuity and/or change and connect that continuity and/or change to a contemporary issue. 4.4 The student will use his/her understanding of continuity and change to construct a model for contemporary reform.
Relationships among people, places, ideas, and environments are dynamic.	5.1 The student will recognize and evaluate dynamic relationships that impact lives in communities, states, and nations. 5.2 The student will analyze the context of significant relationships and draw conclusions about a contemporary world. 5.3 The student will investigate the relationship among people, places, ideas, and/or the environment and connect those relationships to contemporary issues. 5.4 The student will use his/her understanding of these dynamic relationships to create a personal, community, state, and/or national narrative.

II.2. Test Design and Development

The Center for Education Testing and Evaluation (CETE) worked with KSDE to determine the content to be assessed by the KAP tests for each subject area and grade level. The development leading to the 2017 KAP test administration occurred over multiple years. Table II-7 outlines the test-development timeline for the four subjects: ELA, mathematics, science, and HGSS.

Table II-7. Development Timeline for the KAP Assessment

Milestone	Date	Note
ELA/Mathematics		
Adoption of KCCRS	October 2010	
KCCRS item development	2011 to 2016	Determined on a yearly basis.
KCCRS items included in the summative assessment	Spring 2012 to spring 2014	Machine-scored items only. Included to provide schools and districts with a performance snapshot on the KCCRS but not included in accountability measures.
Operational non-adaptive assessment	Spring 2015	Operational items are machine scored only. Performance tasks are field-tested, not machine scored.
Standard setting	Summer 2015	
Operational three-stage adaptive assessment	Spring 2016	Operational items include machine-scored items and hand-scored on-demand writing in ELA and mathematics performance tasks. Includes embedded field-testing for machine-scored items. HGSS MDPTs also contribute to ELA scores.
Operational two-stage adaptive assessment	Spring 2017	Operational items are machine scored only. Includes embedded field-testing for machine-scored items.
Science		
Adoption of KCCRS	June 2013	
KCCRS item development	2015 to 2016	Determined on a yearly basis.
Census field-testing	Spring 2016	Machine-scored items only.
Operational two-stage non-adaptive assessment	Spring 2017	Machine-scored items only.
Standard setting	Summer 2017	
HGSS		
Adoption of KCCRS	April 2013	
KCCRS item development	2012 to 2016	Determined every other year.
Census field-testing	Spring 2015	Both machine-scored and human-scored

Milestone	Date	Note
		(MDPT) items.
Operational non-adaptive assessment	Spring 2016	Both machine-scored and human-scored (MDPT) items. No field tested items.
Standard setting	Spring 2016	

Note. MDPT = multidisciplinary performance task, an on-demand writing task based on primary source documents.

II.2.1. Test blueprints. Test blueprints that specify the number or proportion of items required for each claim (test category: machine scorable or the written portion for HGSS) are presented in Table II-8. ELA and mathematics have the same claim proportions across grades. Science and HGSS percentages vary slightly across grades; the maximum ranges across grades are presented in the same table.

Table II-8. Test Blueprint by Subject and Claim/Category

Claim/ category	Proportion of items by claim or category			
	ELA	Mathematics	Science	HGSS
1	54%–61%	63%–69%	34%–42%	65%–75%
2	39%–46%	11%–13%	26%–38%	25%–35%
3		9%–13%	27%–36%	
4		9%–13%		

The KAP assessment does not specify total score, score point by claim, or proportion by item type in the blueprints because test construction operates under the principle of selecting the most appropriate items for the test. Considerations include items necessarily meeting the content specifications, items having better statistics being selected with a higher probability than items with less desirable psychometric characteristics, and items being developed to ask the question in the most valid manner. The consideration for validity means that items may be technology-enhanced items, multiple-choice items, or performance tasks, depending on the evidence that test takers need to provide to demonstrate mastery. Scores of these items cannot be categorized by item type because some technology-enhanced items have polytomous scores and some performance tasks have dichotomous scores.

Additionally, the blueprints also do not specify the proportions of depth of knowledge (DOK) levels required for the assessment. Because the content standards themselves are organized into an assessment framework of Claims and Targets, each target has a recommended maximum DOK. Items are written to assess at varying levels of cognitive complexity to support the requirement that the test have good measurement characteristics across the range of examinee proficiency. However, the item pool specifications indicate that at least 50% of the items for each target are at the maximum level of cognitive complexity.

II.2.2. Test design. The adoption of new content standards resulted in a transition period for

KAP in 2015 and 2016. This transition process dictated differences in the test design across establishing years. Furthermore, additional changes were made as the testing program evolved. Some of the changes from 2016 to 2017 include the following:

- reduction from a three-stage to a two-stage adaptive design for ELA and mathematics;
- reduction of the overall number of operational items in the ELA test;
- reduction of the number of embedded field test items in ELA and mathematics;
- shift from fixed-form, census field-testing to operational two-stage non-adaptive assessment for science with embedded field-testing;
- standard setting for science in summer 2017; and
- no HGSS administration in 2017 (as per the every-other-year schedule).

In addition, the 2016 ELA test included listening (Claim 3) items, while the 2017 ELA test does not contain listening items and provides only two claims. For the adaptive test in ELA and mathematics, assignment of the Stage 2 block is determined by the ability estimates based on students' answers to Stage 1 items. Because students' abilities are unknown at the beginning of the test, the Stage 1 block is set at a medium difficulty level and includes a wider range of item difficulties to meet the abilities of the majority of students.

For each test, accommodations are provided to students with special needs (see chapter V. Inclusion of All Students). Each stage has a block of items designated for students who need accommodations. When review panels or accessibility experts determine that items are not appropriate for students with special needs, those items are modified. Accommodations are assigned to students who requested them during registration for the KAP assessment.

Table II-9 shows the test design of the KAP assessment, and Table II-10 presents the number of blocks and block difficulty levels for each stage by subject.

Table II-9. Test Design for the KAP Assessment

		No. of items			
Subject	Grade	Total	Stage 1	Stage 2	Note
ELA	3–8, HS	55	30	25	Adaptive
Mathematics	3–8, HS	60	30	30	Adaptive
Science	5	39	20	19	Non-adaptive
	8, HS	44	20	24	Non-adaptive

Note. HS = high school.

Table II-10. Number and Difficulty of Blocks for the KAP Assessment

	No. of blocks		Block difficulty	
Subject	Stage 1	Stage 2	Stage 1	Stage 2
ELA	2	2	medium	easy hard
Mathematics	2	2	medium	easy hard
Science	2	2	medium	medium

Note. For test security reasons, each stage has multiple blocks.

II.2.3. Operational test construction. Domain sampling refers to the selection of a sample of test items from a well-defined population of items (Crocker & Algina, 1986). Student performance on those sampled items is used to infer student proficiency in the tested content area; therefore, the selection of items and item quality will affect the validity and reliability of student ability estimates.

Koretz (2008) noted a few important factors in assuring the generalizability of the test results: motivate respondents in testing, word questions appropriately, and sample items from the domain to achieve content representativeness. Content representativeness is the optimal goal of operational test construction. This goal is achieved by building a test based on a tightly specified test blueprint. However, item quality, both in wording and item statistics, also plays an essential role in test quality and is evaluated in each test construction process. The test construction process is similar across years for both initial test development and ongoing maintenance of the bank of test forms. The process starts with item screening. This involves summarizing item-pool quality from both the content perspective and considering statistical/psychometric aspects to identify eligible items.

- Items and passages are approved by KSDE prior to field-testing and are reviewed by panels of external stakeholders for appropriateness and alignment.
- Following field-testing, items are reviewed for content and psychometric characteristics to rank items for preference in inclusion in assessments.
- Candidate test blocks are assembled following the content specifications in the blueprint and preferentially selecting items with the best psychometric characteristics (e.g., higher slope item preferred over a lower slope when the content characteristics are otherwise parallel).
- Candidate test blocks are reviewed to eliminate item enemies (e.g., items that might clue answers to other items).
- Final test blocks are submitted to psychometrics to confirm the psychometric properties and, for ELA and mathematics, to confirm the adequacy of item selection in the routing stage (first block).

Each subject has additional guidelines for test construction.

II.2.3.1. ELA and mathematics test construction guidelines.

- Different test forms include approximately the same number of items per claim.
- Different test forms include approximately the same number of items per target.
- Different test forms include approximately the same number of items per DOK.
- ELA and mathematics Stage 1 block includes a wide range of item difficulties, and their average difficulty is of moderate level.
- Linking items should have robust item statistics and match the test blueprint. They are placed in Stage 1.
- In mathematics, the Stage 1 items are ordered from easiest to hardest within each claim.
- In ELA, passage-based items are ordered according to established protocol (i.e., starting with main idea and followed by specifics) and referencing the order of appearance in the text.
- In ELA, Claim 2 items are generally ordered from least difficult to most difficult.
- Embedded field test items are included in Stage 1.

II.2.3.2. Science test construction guidelines.

- Different test forms use approximately the same number of items per claim.
- Different test forms use approximately the same number of items per target.
- Different test forms use approximately the same number of items per DOK.
- Each block includes a wide range of item difficulties, and their average difficulty is of moderate level.
- Science items are ordered by difficulty within claim. However, different forms may have a different order of claims. For example, one form in Stage 1 begins with physical science, yet another form in Stage 1 begins with life science. The items are parallel in claim, target, difficulty, and DOK but are reordered for test security purposes.

II.2.4. Item pool evaluation. Because both ELA and mathematics use multistage adaptive tests, more items are consumed. The number of quality items in the item pool is essential to the success of the design. This section addresses item-pool quality from three perspectives: content alignment, item count by content standards, and simulation of paths to different blocks in Stage 2.

II.2.4.1. Alignment study of adaptive test item pool. In fall 2014, edCount and CETE drafted a plan to investigate multiple facets of KAP items: the use of items, gaps between the expected and actual use of items, and alignment throughout the test-development process. After a multiyear effort, the *Kansas Assessment Program Alignment Evaluation Report 2015–2016* (the *Alignment Study* hereafter) was completed in July 2016.

The *Alignment Study* gathered a wide range of evidence to address the quality of items and performance tasks in association with test blueprints. The evidence included the following:

> item quality, alignment, coherence, and accessibility; blueprint quality and alignment; and test form alignment to targets and intended blueprints. This enhances item reviews of 2014–15 through the inclusion of additional items for review and the addition of blueprint and test-level reviews, which provide evidence regarding the degree of alignment between the

assessments and the claims and targets. (Forte et al., 2016, p. 1)

Unlike typical alignment studies that are designed for post hoc evaluation, edCount used Forte's (2013, 2016) framework. A process was developed by edCount to include items in development and recently field tested items in the early evaluation stage and emphasized the alignment between item, blueprint, and content standards. This first phase of the study was to ensure the content adequacy and alignment of the item pool. Then, in spring 2016, approximately 355 ELA and 234 mathematics items of the 2016 KAP operational test were reviewed by panels of content experts, who were instructed to evaluate whether each item contributed to the overall blueprint and to determine that the test forms matched the intent of the assessment as laid out in the test documentation. This section summarizes the results of the second phase of the alignment study.

The edCount blueprint review panel, comprising four internal content and research staff members, used the internally developed protocols to assess the connections among the KAP KCCRS, the test blueprint, and item-bank metadata. The panel concluded that the item pool for all grades of both ELA and mathematics met the following requirements: at least six items addressed each claim on the blueprint, at least one item slot in the blueprint was assigned for each target in the content emphasis document, and the percentage of items addressing each claim met expectations. Because KAP did not have DOK blueprints, averages of DOK by target were computed. ELA DOK was 2.4 for all grades; mathematics DOK ranged from 2.4 to 2.6. The values indicated that there were more items in the Level 3 DOK (higher cognitive complexity). Additionally, evaluation of operational pathways and items indicated that each pathway adhered to the blueprints, and operational items reflected the breadth and depth of the KCCRS.

When selecting items for the adaptive test, CETE uses the stage approach in that the test blueprint divided into stages and blocks within each stage are parallel forms. Thus, all possible combination of adaptive routings will yield test forms that match the blueprint.

II.2.4.2. Item count by content standard. Table II-11 presents the number of items and proportion by claim in the overall ELA and mathematics operational test panels for the 2017 test construction. Each student took a subset of these items as individual students were routed through the test from Stage 1 into Stage 2–Hard or Stage2–Easy. The proportions of ELA and mathematics items by claim in the total operational test panel align with their test blueprints.

Table II-11. Percentages of Items by Claims

Grade	No. of ELA items	ELA claims (%)		No. of math items	Mathematics claims (%)			
		1	2		1	2	3	4
3	86	59	41	88	65	11	13	11
4	82	57	43	80	66	13	10	11
5	81	57	43	87	63	13	11	13
6	85	56	44	85	69	12	9	9
7	74	61	39	80	66	11	11	11
8	80	56	44	83	63	11	10	11
10	68	54	46	83	66	11	11	12

II.2.4.3. Item statistics. The ELA and mathematics tests use a 1-2 design and yield two possible pathways. Block difficulty levels and possible pathways are presented in Table II-12.

Table II-12. Pathways of Multistage Design for ELA and Mathematics

Pathway	Stage	
	1	2
1	Medium	Easy
2	Medium	Hard

Simulations for the multistage adaptive test were performed during test construction. As mentioned earlier, Stage 2 block assignments depend on students' performance in the previous stage. The algorithm used to determine block assignment is the test information function (block information, in this case) of item response theory (IRT). The block that provides the most test information is selected for administration. For example, Figure II-1 shows ELA grade 3 Stage 2 block information-function curves. It can be seen that the easy block has more information function in the theta range of (−4.0, −0.3), while the hard block has more information in the range of (−0.3, 4.0). If a student's Stage 1 theta estimate is 0.4, the hard block will be administered. The results of pathway routing for ELA and mathematics can be found in Appendix E, Path Reliability.

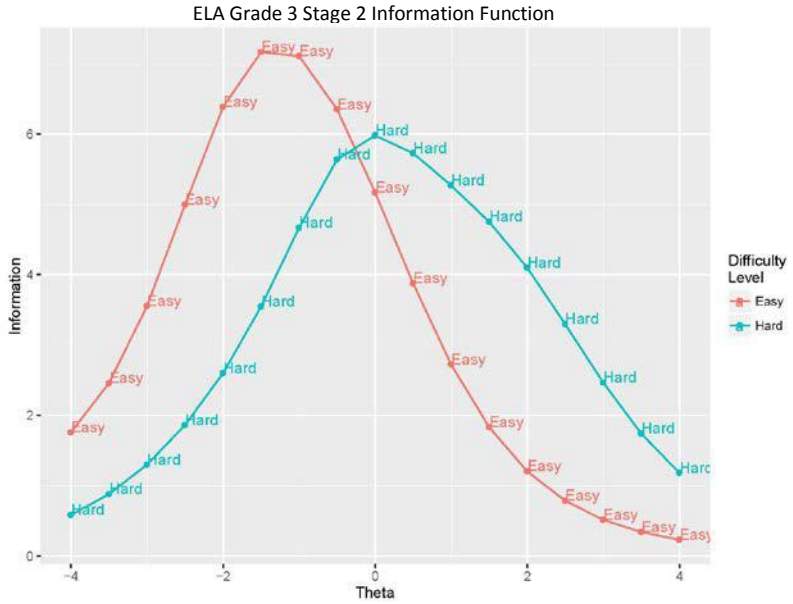


Figure II-1. An example of Stage 2 block information curves.

II.3. Item Development

Item development entails various efforts to ensure item quality, including ongoing research into best practices and new item types, developing and using subject-area item specifications, updating materials for item-writer training, recruiting new or additional item writers, conducting item-writer training for new item writers or a refresher training for continuing item writers, creating items, and reviewing and revising items. Item review is conducted in two phases: first, when items are created and next, after items are field-tested. In the first phase, both CETE content experts and trained, external item reviewers review items. Before appearing on any assessment, items are reviewed by content reviewers, bias and sensitivity reviewers, and KSDE staff. CETE staff use item-review feedback to revise test items as needed. Items are then prepared for field testing, according to test specifications and following established guidelines for both general and accommodated presentation. After field testing, item and test data are analyzed; this data analysis guides decisions about the use of items on future assessments. The following section describes typical procedures for different stages of item development.

II.3.1. Passage selection and review. For ELA, the process starts with the identification of appropriate public domain works or the commissioning of passages as work-for-hire. CETE's passage development team has built a strong network of both regional and national authors, allowing the team to generate high-quality, original passages capable of supporting item development.

The ELA team uses several resources, both qualitative and quantitative, to analyze text complexity and to guide grade-level placement. Assessment passages include commissioned, permissioned, and public domain readings. Passages from all sources undergo multiple rounds of internal review, including editorial, content, bias, sensitivity, and accessibility reviews. For

example, CETE content specialists and accessibility specialists review passages for accessibility issues and content accuracy (e.g., inaccurate or outdated science information). Outside experts with knowledge of low-incidence disabilities or with specific subject-matter expertise are also consulted as needed.

Passages that are accepted at the internal review are then reviewed by an external panel of educators. These external passage-review panels are formed by grade band: grades 3–5, grades 6–8, and high school. Each panel includes educators with backgrounds in EL and special education. Reviewers are provided with training and detailed instructions regarding how to review passages through CETE's secure, online reviewing system. Panelists then review passages at their own pace and provide feedback and placement recommendations by a given deadline.

Passage reviewers, during both internal and external review, use rubrics of both qualitative and quantitative measures to examine text complexity and grade-level suitability through text structure, language features, and knowledge demands. CETE uses the Flesch-Kincaid score as a quantitative measure for longer passages. However, passages of only 350–450 words are not long enough to give an accurate Flesch-Kincaid reading, and most measures of text complexity are inadequate for the task of analyzing poetry. In those cases, CETE considers sentence length and complexity to gauge an initial grade placement. Qualitatively, CETE looks at each set for vocabulary, knowledge demands, topic familiarity, and interest level.

In addition, both internal and external reviewers consider the following passage components.

- Length: Are the texts of reasonable length for students? Are the texts long enough and rich enough to support all or most of the item content established in the item and test specifications?
- Bias or sensitivity: Are all groups portrayed accurately and fairly? Does the passage demonstrate awareness of different cultures and sensitive topics in the state (e.g., natural disasters, politics)?
- Overexposure: Is the passage already commonly taught in the school or district, or is it used frequently in anthologies or lesson plans?
- Interest level: Will more than half of students be at least moderately interested in the passage?
- Images: Are there any concerns related to the accessibility or content of images? Should images be added to enhance or support the passage?
- Prior knowledge: Should introductions be included to provide historical context or background information?

Reviewers are then asked to recommend a grade for each passage based on complexity and other considerations. After the passage-review window is completed, reviewers are invited to an optional telephone discussion of the passages. After compiling the information and summarizing the overall data collected from the review, CETE shares the results and passages with KSDE for

approval of grade placement. Based on item-pool needs (e.g., complexity levels, text types, topics), some passages are selected for item development. Remaining passages are held for future development.

II.3.2. Item writers. Some item-development staff are full-time employees of CETE. Other item writers have been University of Kansas (KU) graduate research assistants (GRAs). The GRAs are recruited and hired based on their training in a given subject, prior item-writing or test-development experience, or previous teaching. Because ELA, mathematics, science, and HGSS tests cover a wide range of knowledge and skills and also incorporate diverse real-life topics as item contexts, GRAs who write items for the assessments have come from a variety of academic areas, including curriculum and teaching, English, mathematics, economics, pre-med, classical languages, biology, computer science, physics, social welfare, and educational psychology. Additionally, panels of educators and subject-area experts from outside of CETE have assisted in the development of specialized, open-ended items, including extended writing tasks in multiple content areas in HGSS and ELA, and mathematics performance tasks.

II.3.3. Item-writing training. Before writing items for the KAP assessment, item writers are trained in the use of KAP subject-area item specifications in the writing and reviewing of items. All item writers receive training in several topics, including the following:

- the KCCRS,
- validity and reliability,
- alignment,
- differentiating between cognitive complexity and difficulty,
- evidence-centered design,
- principles of universal design (UD) and accessibility,
- bias and sensitivity, and
- item types.

To guide the item-writing process, item writers are trained in content, format, structure, stem structure, answer choice development, accessibility, bias and sensitivity, and traditional and nontraditional item types. Besides learning fundamental principles of item writing, item writers also receive training in item review so they can objectively evaluate their own products as well as others' items. Key points of these guidelines are presented below.

II.3.3.1. General guidelines.

- Write items that have clearly correct answer choice(s), with other answer choices clearly incorrect.
- 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.

II.3.3.2. 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.

II.3.3.3. Format guidelines.

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

II.3.3.4. Structure guidelines.

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

II.3.3.5. 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., "best," "most likely").

II.3.3.6. Answer-choice development guidelines.

- Order answer choices logically.
- Create independent answer choices that do not overlap.
- Write answer choices that are roughly of the same length and parallel in structure.
- Do not offer "all of 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 such as "always" or "never."
- Create plausible distractors.
- Create distractors that take advantage of common errors and misconceptions.
- Answer keys should be roughly uniform in distribution.

II.3.3.7. Accessibility guidelines.

- Consider the access needs of special populations and the ways in which accommodations affect an item's intent.
- Use simple sentence structures.
- Minimize the 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.

II.3.3.8. 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 religious references, such as holidays.
- Ensure that items are not related to socioeconomic status or family attributes.
- Use artwork that reflects the diversity of the student population.

Item-writing training also includes extensive practice. Participants discuss DOK for specific standards, examine practice items for alignment to content standards, and determine whether practice items are written to the appropriate difficulty level. Participants also practice writing items and receive feedback from CETE staff. Additional specialized training for performance tasks is provided to content experts.

II.3.4. Item writing. Traditionally, the internal item writing and review process for all content areas starts among the item writers. Because of the necessary research to make sure that the context is technically correct, the initial item writing can take anywhere from a few hours to a few days. The item writer has to match the item to the metadata requirements and ensure that the item follows the rules of item writing, that the content is correct and that any surrounding context is accurate, that the language is appropriate for the grade being tested, and that the correct answer(s) is (are) correct and verified.

The item writer will send a completed item or a set of items (particularly for ELA for passage-based items) to a fellow item writer to review. They discuss the items, the alignment to the standards, and the cognitive complexity demands, and the item writer revises any items as needed. The items are then passed onto a content specialist or test-development assistant for further review.

Following the content specialists' review, the item is passed to editing. If graphics are needed, a content specialist will provide instructions to the graphic artist regarding the rendering of the stimulus and then will confirm that the completed graphic meets the intended function within the item. When the editors have finished editing the items, the content specialists re-review the items prior to passing the set onto the content lead and psychometricians for adherence to best item-writing practices.

The content lead either approves the item (and graphics if needed), makes his or her own edits, or sends it back to the item writer, content specialist, or graphic artist for revisions. Items are then reviewed by a psychometrician for adherence to best item-writing practices; they are often reviewed often simultaneously by an accessibility expert for adherence to principles of UD and issues in accessing the item that may be encountered by students with disabilities or students who are EL. The accessibility reviewer may refer items to further review by experts with knowledge in low-incidence disabilities, such as blind/low vision or deaf/hard of hearing. Following these reviews, an item may be returned to the editing team if substantial changes have been made. After the completion of internal reviews, items are sent to external committees and KSDE for review.

CETE relies on teacher expertise throughout the task development process for performance tasks. For example, HGSS content experts identify an array of primary and secondary source documents and create sets of documents (or excerpts) from these sources to serve as the basis for both the items about the primary source and an on-demand writing task involving analysis of the documents. With guidance from CETE and KSDE staff, the external content experts develop Document Focus questions for each document and develop a series of writing prompts for pairs or trios of documents to be analyzed together.

II.3.5. Item reviewers. The item-review process involves several stages.

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

Much of the internal item review process was described with item writing. This section primarily describes the external reviewers and review processes.

CETE content experts and KSDE staff recruit item reviewers from Kansas educators for two separate types of reviews: content review and bias and sensitivity review. Prospective item reviewers complete an online survey in which they indicate their demographic information, teaching experience, professional qualifications, content expertise, experience with the standards, and special education or EL endorsements or training.

Content-review panels for ELA and mathematics are typically formed by grade band: grades 3–5, grades 6–8, and high school. Content-review panels for HGSS and science are formed by grade, but some reviewers serve on more than one panel because domain content knowledge often extends above or below grade levels. Bias and sensitivity panels are assembled and include members of various groups to reflect the diversity of Kansas. Similar to the passage review

process, item reviews are processed through a secure, online reviewing system. After completing a web-based training session, reviewers evaluate items at their own pace and provide feedback by a given deadline.

II.3.6. Item review. All item reviewers must complete two web-based sessions of item-review training: completing the online review system and completing the specialized training for either bias and sensitivity training or content-review training. The training sessions include information about the KSDE–CETE partnership, test and item security, item-writing guidelines, and the item-review process. Item-review training also provides participants with practice items and CETE staff contact information.

Bias and sensitivity reviewers are asked to identify barriers that may prevent students from demonstrating what they know and are able to do when those barriers are not related to the content standards. These barriers may include unfamiliar or variably familiar language; linguistic complexity; potentially sensitive topics; presentation of stereotypes, including emotions, regions, or occupations; accessibility for special populations; and issues with cultural or prior knowledge. The reviewers are given a code sheet that provides code categories and descriptions for possible concerns. When reviewers flag items for bias or sensitivity concerns, they use these codes and can provide additional details in a comment section. A code is also assigned to indicate there is no barrier, bias, sensitivity, or other concerns for clarity and record-keeping purposes. Descriptions of concerns 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

Content reviewers verify alignment to the content standards and the appropriate DOK; they judge the appropriateness of the item, including its content, context, and vocabulary for the grade and subject; they check the correct answer and evaluate the extent to which the incorrect answers would give them useful information about what their students do not know; they evaluate the need for any included graphic or stimulus, and if one is included, they comment on its utility and clarity; and they identify any possible concerns about accessibility. Content reviewers also attend to the alignment of items to assessment targets, checking that items adequately address part of the target and elicit evidence for at least part of one evidence statement. In general, content reviewers check items for the following:

- appropriate, grade-level vocabulary;
- a clear, complete statement or question;
- grammatically correct text;
- a correct key;
- accurate, relevant graphics; and

- well-designed answer choices that do not require background knowledge outside of the content area and that are 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 can recommend that items be accepted, revised, or rejected, and give specific reasons for their decisions (e.g., "item aligns better to this assessment target").

II.3.7. Universal design (UD) in test development. UD in item and test development not only allows for the participation of the widest range of students, but it also bolsters the validity of score inferences. KAP's comprehensive inclusion rules mean that KAP tests include virtually all Kansas students. While the initial intention is to meet the interests of special-needs students, the benefits of universally designed assessments should apply to all students with diverse characteristics.

Item-writer training teaches participants about UD concepts, including a definition of UD and examples of test items that adhere to UD principles. Additionally, the item-writer guidelines include many UD principles. The following are some focuses of UD in KAP's development.

- Item writers are trained to become aware of and sensitive to issues of cultural and regional diversity.
- Both internal and external reviewers of items and test specifications strive to ensure that no barriers stem from a lack of sensitivity to ability, culture, or other characteristics.
- The tests are compatible with many accommodations and a variety of widely used adaptive equipment and assistive technology without changing the meaning or difficulty of test items.
- The language used in test materials is direct and concise. Additionally, unnecessary images and text are omitted to avoid distracting students.

II.3.8. Field testing. In general, field testing of new KAP items uses the embedded-model approach. The main advantage of an embedded field test is that the examinees cannot differentiate items that count toward their score from field test items, thereby using the same care to answer the field-test items. This trait improves the field test item data quality and provides more robust item-parameter estimates.

II.3.9. Field test data analysis. Field test item analyses include classical item analysis, IRT calibration, model fit evaluation, and differential item functioning (DIF) analysis. Items that are too easy or too difficult, that do not discriminate students' ability well, or that have large DIF are flagged according to predetermined criteria. The statistics and flags are added to the item pool for use in test construction. Note that because this report focuses on the technical characteristics of the operational tests, field test statistics are not presented.

II.3.10. Data review. Following field test item analyses and prior to test construction, the

content team reviews item statistics. Items with statistical flags are used only when the item pool does not have other items for blueprint coverage. When flagged items are used, they undergo extra review and discussions.

II.4. Test Administration

Large-scale assessment requires a standardized test-administration process to prevent the unintended effects of administration differences. The standardized test-administration procedures are described in the *Examiner's Manual 2016–2017* and *Tools and Accommodations for the Kansas Assessment Program (KAP) 2017* (hereafter *Tools and Accommodations*). The *Examiner's Manual* provides information regarding standardized test administration for districts, schools, and teachers; *Tools and Accommodations* provides guidance regarding the use of available accessibility tools and features for assessments. Teachers who administer the KAP assessment are required to sign an agreement to follow the guidelines and to show that they have learned about test security and ethical test practices.

In the *Examiner's Manual*, test security procedures are described in multiple sections. Test Security Plan and Test Security Guidelines are found in Section 2: Test Coordinators. Test Security and Administration is in Section 4: Teachers. The Test Security Guidelines section of the *Examiner's Manual* explicitly explains to the District Coordinator the test security practices and actions after detected test security breach, loss of materials, or any other deviation.

II.4.1. Test administration and security training. All Kansas District Coordinators must take the test administration and security training during the preconference at the KSDE Annual Conference in October or with online training materials available from the KSDE assessment website. District Coordinators will train building-level personnel before the local test. All local personnel administering state assessments must read the *Examiner's Manual* and sign an agreement to abide by state ethical testing practices. See Appendix A for the training PowerPoint.

II.4.2. Monitoring test administration. District and Building Test Coordinators can monitor student test progress via the Kansas Interactive Testing Engine (KITE®) Educator Portal. This process is described in Section 5: Test Administration of the *Examiner's Manual*.

During the testing window, KSDE staff and members of the Kansas Assessment Advisory Council visit a random sample of Kansas schools to monitor administration and test security. The *State Monitor Quality Assurance Checklist for Test Security and Ethics* is posted on the KSDE website, along with other assessment-related documents and resources to assist districts and schools in understanding the KAP administration.

Provision of accommodations is handled in two ways: by test administrators and by the online test portal. Information about accommodations handled by test administrators are not available in the online system. Accommodations built into the online testing portal are discussed in section V.3 Accommodations. Evaluation of the consistency between the accommodations included in

the individualized education program (IEP) and during the assessment cannot be conducted because the IEP information is not available. However, this is part of the state quality monitoring process conducted by KSDE.

II.5. Systems for Protecting Data Integrity and Privacy

The electronic item bank, online administration system, and student responses are stored in KITE Suite, which is designed and maintained by the Agile Technology Solutions (ATS) center, part of the Achievement & Assessment Institute (AAI) at KU. Multiple portals are designed within KITE Suite to serve the needs for item and test development (i.e., Content Builder), for educators to input and access test and student information (i.e., Educator Portal), and for online testing (i.e., Kite Client).

AAI fully understands the importance of test security in both protecting student information and ensuring valid interpretation of test data. The physical security requirements are met by using hosting providers that conform to SAS 70 auditing standards for physical access and PCI compliance. Most of the project management, test development, and data analysis activities take place at CETE. CETE's on-campus offices are in a secure wing that can be accessed only with a key. ATS's off-campus offices are accessible only with an electronic key card. In general, most work is done at one of our sites using secure server systems. CETE and ATS staff access those servers via a secure VPN connection when they need to work remotely.

All KITE applications handle educator and administrative passwords using industry-standard encryption techniques; users must create strong passwords and may change their own passwords at any time in accordance with the password policy. All applications generate access records that can be reviewed by system administrators to track access. All released items exist in a separate pool from items used for summative purposes, ensuring that no items are shared among secure and nonsecure pools. Only authorized users of the KITE assessment system have access to view items.

In accordance with FERPA, students', teachers', operators', and administrators' access to personal student data is limited to student records in which that person has a legitimate educational interest. All users are provided the minimum amount of necessary access. Throughout each school year, security levels, groups, and access are reviewed periodically to ensure continued compliance.

Operational access to all servers is controlled by keys that are provided only to system administrators who manage the production data center in the operations team. Access to the networking equipment and hardware consoles is limited to the data center itself; remote access to these devices is limited to the data center–specific administration host.

Access to individual KITE applications is controlled according to the policies set forward for that application and the data the application maintains. All access policies and accounts are reviewed periodically to ensure that access to systems is limited to the appropriate populations.

In addition to physical and electronic security measures, test security is promoted through required training and certification requirements for test administrators. Test administrators are expected to deliver assessments with integrity and to maintain the security of assessments. State, district, and school users are expected to complete the security agreement within Educator Portal each year. By accepting the security agreement, users agree not to store or save assessment materials to computers or personal storage devices, to not print assessment materials, and to not share personal passwords with others.

III. Technical Quality—Validity

As defined in the *Standards for Educational and Psychological Testing* (the *Standards* hereafter), validity refers to "the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (American Psychological Association, American Educational Research Association [AERA], & National Council on Measurement in Education, 2014, p. 11).

The *Standards* (American Psychological Association et al., 2014) provide a framework for describing the sources of evidence that should be considered when evaluating test score validity. These sources include evidence based on (a) test content, (b) response processes, (c) internal test structure, (d) relationships between test scores and other variables, and (e) consequences of testing. Other sources of evidence also can bolster the validity argument. For example, when IRT is used to analyze 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. This technical manual mainly describes aspects of the KAP assessments that support KAP test score interpretations and uses.

III.1. Overall Validity, Including Validity Based on Content

Because the intended uses of the test scores are one source of evidence in a validity study, the purposes of the test should be identified before providing evidence to support test validity. The purposes of the KAP assessment, described at the beginning of this manual, include (a) measuring specific claims related to the KCCRS, (b) providing Annual Measureable Objectives for state accreditation, (c) reporting student's academic performances, and (d) using with local assessment scores to assist in improving educational programs in the four subject areas.

Evidence gathered on content validity, alignment, cognitive process, and internal structure supports the use of the KAP assessment to measure the KCCRS content as defined in the test blueprints. Information on test reliability, fairness and accessibility, and scoring and scaling justify the use of KAP test scores for Annual Measureable Objectives and reporting student's academic performances. Validity evidence from other sources, such as using KAP scores to predict ACT scores, uses additional data to assist educators.

III.1.1. Content validity. Evidence of content validity for the KAP assessment comes from the alignment between KAP items and the KCCRS and the congruence between the test and test blueprint. The following procedural steps are used to evaluate the content validity of the KAP assessment.

- Evaluate the alignment between KAP items and KCCRS.
- Evaluate the degree to which the KAP test blueprint represents and aligns with the knowledge and skills described in the KCCRS.
- Conduct content reviews of KAP items using a panel of content experts to see whether

the items measure the intended construct or whether sources of construct-irrelevant variance exist.
- Conduct fairness reviews of KAP items to avoid bias and sensitivity issues related to specific subpopulations.

The first two chapters of this technical manual present validity evidence related to test development and a summary of the alignment study. As described in those chapters, all KAP items are developed and aligned with the KCCRS, and item development followed well-established procedures. After items are developed, they undergo multiple rounds of content and bias reviews. After field test administration, items' statistical properties are reviewed. Items are evaluated by content, psychometric, and KSDE reviewers before selection for operational use. Tests are also administered according to standardized procedures, with accommodations for students with special needs. Specific efforts to ensure content validity are summarized below.

- Webb's (1997) DOK model is used to identify the cognitive complexity of KAP items, ensuring that items cover the range of cognitive complexity. Although DOK distribution is not specified in the test blueprint, item specification documents provide information on the expected DOK for each assessment target. Item writers use this information to write items that match the DOK expectation of each assessment target. The analyses of DOK distributions by subject and grade are presented in Tables IV-7 through IV-9.
- Qualified item writers are selected and trained to ensure they write high-quality items.
- Detailed item- and passage-development guidelines are established and used to train item writers, who also participate in guided item writing.
- CETE content specialists and editors review each new item to make sure all items align with the KCCRS; they also consider grade-level appropriateness, DOK, graphics, grammar and punctuation, language demand, and distractor reasonableness.

- Content committees comprising Kansas educators then review items and consider, among other elements,
    - overall quality and clarity,
    - KCCRS alignment,
    - grade-level appropriateness,
    - difficulty level,
    - DOK,
    - appropriate sources of challenge (e.g., item difficulty is not related to unintended content or skills),
    - answer correctness,
    - quality of distractors,
    - graphics,
    - appropriate language demand, and
    - absence of bias.
- An external bias, fairness, and sensitivity committee reviews items for issues related to diversity, gender, and other factors.
- Before items are selected for operational use, several statistical analyses are conducted, including classical item analysis, distractor analysis, and DIF analysis. CETE staff again carefully review items' statistical characteristics.
- Administration of KAP assessments is standardized and includes accommodations. Students are given ample time to complete the tests to avoid speediness issues).
- The item-pool analyses described in section II.2.4.2 Item Count by Content Standards of this manual show that each claim has an adequate number of items to cover test blueprints for all subjects and grades.

III.2. Validity Based on Cognitive Process

Response-process evidence examines the extent to which the cognitive skills and processes students use to answer an item match those targeted by item writers. While studies that investigate students' cognitive processes, such as think-aloud, are not planned, alternative evidence is established during the item-development process and with the development of performance level descriptors (PLDs).

During the item-development process, items were written by content experts who have been trained on proper item-writing approaches. Then items were reviewed by content experts who had direct experience with students. The content standards provide content specification and imply a target DOK. The DOK component guided item writers to use language that elicits the cognitive process required by the content standards and guided item reviewers to evaluate the cognitive process required by items.

The PLDs are also used to reflect the cognitive process required for the specific content area. For example, the PLDs of grade 5 science presented in Appendix H provide introductory policy statements for each performance level in the table. The policy statements are extracted from the table and are listed below.

- Level 1: Students show a limited ability to understand and use the science skills and knowledge needed for college and career readiness.
- Level 2: Students show a basic ability to understand and use the science skills and knowledge needed for college and career readiness.
- Level 3: Students show an effective ability to understand and use the science skills and knowledge needed for college and career readiness.
- Level 4: Students show an excellent ability to understand and use the science skills and knowledge needed for college and career readiness.

As performance levels increase, the expectations of students' proficiency or cognitive process increase. As shown above, from Levels 1 to 4, the required ability of students to understand and use the science skills and knowledge change from "limited" to "basic," then to "effective" and, finally, to "excellent." The PLDs were written and reviewed by content experts and educators.

III.3. Validity Based on Internal Structure

As described in the *Standards* (American Psychological Association et al., 2014), internal-structure evidence refers to "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (p. 13). For each KAP assessment, one total test score and several subscores are reported. Multiple sources providing internal-structure evidence relating to the use of both types of scores are discussed below.

III.3.1. Internal construct. Item-test correlations (indicators of item discrimination) are reviewed in this manual in section IV.3.1 Classical Item Statistics. The range of acceptable correlations for adaptive tests is broader than nonadaptive tests because extremely easy and difficult items are included to provide better theta estimates on the two ends of the scale. These extreme items tend to have low discrimination values. Items having negative item-test correlations are excluded; generally, other items should have an item-total correlation of at least 0.250, although lower discriminating items may be used to match the content requirements. As noted, any time an item with less than desirable characteristics is used, the item undergoes an additional layer of scrutiny prior to inclusion. The summary of item discrimination presented in section IV.3.1 shows that all but one item used operationally have positive correlations. The median item discrimination across all three subjects (ELA, mathematics, and science) is in the 0.30–0.40 range.

III.3.2. IRT and model assumptions. The KAP items are analyzed using IRT. IRT is an industry standard for item analysis in large-scale K–12 assessment programs because of its item and person invariance claims. However, it has several model assumptions that need to be fulfilled: model fit, unidimensionality, and local independence. The resulting inferences from any application of IRT depend on the degree to which the underlying assumptions are met.

The current section introduces the IRT models and calibration procedures used for ELA, mathematics, and science. Evaluation of IRT assumptions is presented as evidence of the appropriateness of model selection and is part of score validity.

III.3.2.1. Samples. The 2017 KAP ELA and mathematics assessments use a two-stage adaptive design. Two pathways are designed for the multistage adaptive design (see chapter II. Assessment System Operations). Each student takes a total of 55 ELA items and 60 mathematics items regardless of the pathway. Test blocks are pre-equated; item parameters obtained prior to the current administration are used to estimate the thetas of this year's test. This method does not require re-estimation of item parameters using the current year's data; thus, equating is done prior to test administration, hence the term pre-equated. No additional calibrations were conducted in 2017 for operational ELA and mathematics. Interested readers may refer to the 2016 and 2015 technical reports for further information on the IRT model assumptions.

Science items were administered as a fully operational test for the first time in 2017. These items did not have pre-existing parameters, so a post-equating design (i.e., equating done after the test administration data become available) was employed to estimate science item parameters using 2017 test administration data. A single-group concurrent calibration was used to place all item parameters onto the same scale. To accomplish this, all operational items of the same subject and grade were compiled into one file to create a student-by-item data matrix, which was then analyzed using flexMIRT Version 2.80 (Cai, 2013) for concurrent calibration.

The student data file was cleaned prior to calibration and equating. For example, the estimation sample includes all students who completed the test, except students who needed certain accommodations. Since each subject and grade is calibrated with a single-group concurrent method, the sample size for concurrent calibration equals the number of valid cases. Table III-1 provides numbers of students by subject and grade.

Table III-1. Sample Size for Concurrent Calibration by Grade for Science

Grade	Sample size
5	33,156
8	33,458
11	32,210

III.3.2.2. Missing data. Missing responses require special attention because the coding of missing data can affect item-parameter estimates. There are two types of missing responses: omitted and not administered. Omitted items appeared on the test, but students did not answer them; thus, they are scored as incorrect answers (coded as 0). Not-administered items did not appear on the test form students took but did appear on other test forms and, therefore, are coded as missing.

III.3.2.3. Excluded items. No science items were excluded during IRT calibration.

III.3.2.4. IRT models. The two-parameter logistic (2PL) model (Birnbaum, 1968) and the graded response model (GRM; Samejima, 1969) were applied to dichotomous and polytomous scored items, respectively. The choice of these two models contributes to the consistent and coherent interpretation of item parameters, as the 2PL is a special case of GRM that handles dichotomous items. The 2PL model defines the probability that a student of proficiency θ will answer item i correctly (u) as

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

where a_i is the discrimination parameter and b_i is the difficulty parameter. Discrimination, or differentiation, indicates how well the item distinguishes between students with higher or lower levels of proficiency; difficulty is the degree of item difficulty and is on the same scale as theta.

Under the GRM, the probability that u_i is equal to any observed score category v equals the cumulative probability of scores 0 to $v-1$, minus the cumulative probability of scores v to the maximum score. The probability that the score is v or higher is

$$P(u_i = v|\theta) = \frac{e^{[a_i(\theta - b_{iv})]}}{1 + e^{[a_i(\theta - b_{iv})]}}, \quad \text{(III-2)}$$

where a_i is the discrimination parameter and b_{iv} is the difficulty parameter for score category v. One discrimination parameter is estimated for each item; this parameter may be interpreted as the strength of association between the item and theta. For m response categories, there are $m-1$ GRM b parameters. The b for category v is interpreted as the point on theta where the probability of scoring in category v or higher is 0.5.

III.3.2.5. Evaluating IRT assumptions. The validity inferences from the IRT results depend on the degree to which assumptions of the models are met and on how well the models fit the data. In this section, the assumptions about IRT model fit, unidimensionality, local independence, and item-parameter invariance are evaluated.

III.3.2.5.1. IRT model fit. 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, Table III-2 presents the number of items, the number of misfit items, and the percentage of misfit items for science.

Table III-2. Science Misfit Results by Grade

Grade	No. of items	No. of misfit items	% of misfit items
5	42	0	0
8	48	0	0
11	47	4	9

III.3.2.5.2. 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 (Rosseel, 2012). The one-factor CFA model was considered to fit well if the comparative fit index (CFI) and Tucker Lewis Index (TLI) were 0.95 or greater and the Root Mean Square Error of Approximation (RMSEA) was 0.05 or smaller.

Overall, the science tests in grades 5, 8, and 11, both the CFI and the TLI are around 0.99 and the RMSEA range from 0.01 to 0.03. All the tests may be reasonably treated as unidimensional.

III.3.2.5.3. Local independence. The assumption of local independence means that the response to an item is not affected by responses to other items. This definition is necessary because it secures the foundation of the IRT model: The probability of answering an item correctly is affected only by the item's characteristics and student proficiency. If other items affect an item's response, then the IRT model cannot be used because it fails to incorporate the effects of other items. Local independence is violated when the student responses to items in the latter positions of the test depend on the student responses to their predecessors. In this case, when the first item of the group is answered incorrectly, it will cause the answers to the remaining items to be incorrect. Another, more subtle violation of local independence is when either the question itself, or one of the answer choices, provides cluing that changes the probability of correctly responding to another question.

Evaluation of local independence starts during item development. As long as all test items are written so that they do not depend on the responses to other items, local independence is assured. During test construction, all items on a test are reviewed to ensure neither the items nor the answers clue students to other items on that test.

III.3.2.5.4. Invariance. IRT models claim that item-parameter estimates are invariant up to a linear transformation for all examinees. Bivariate scatter plots and Pearson product-moment correlations were used to evaluate the relationship between the item parameters estimated from subgroups that are expected to have the same ability distributions. To avoid statistical bias caused by outliers, any items with discrimination parameters smaller than 0 or greater than 4, or with difficulty parameters greater than |6|, are excluded from the comparison. The invariance assumption is met if the estimated item parameters for female and male samples are highly correlated.

Here, the subgroups are determined by gender. The scatter plots presented in Figures III-1 and III-2 indicate that science items have strong linear relationships between item-parameter estimates for female and male samples. The items with large discrepancies suggest potential gender DIF. Table III-3 shows that all the Pearson correlations are above 0.90. These results strongly support the invariance assumption for KAP science, especially for item-difficulty parameters.

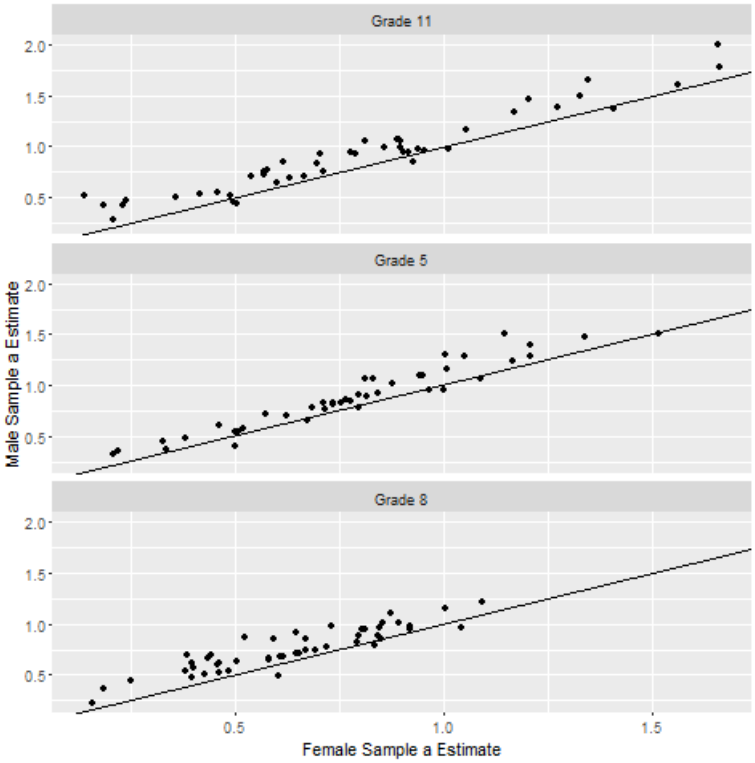


Figure III-1. Science item-discrimination parameter scatter plot by grade.

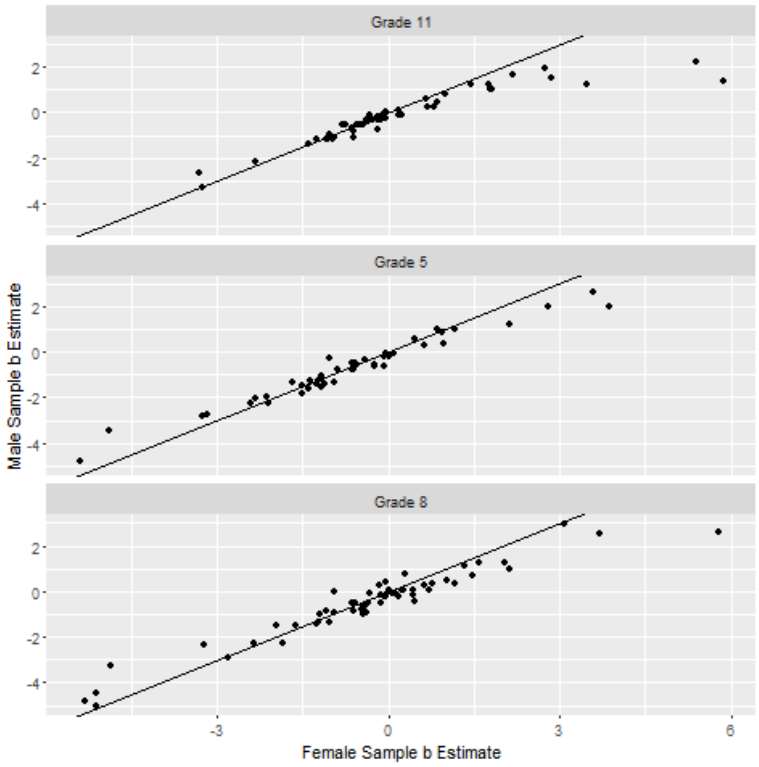


Figure III-2. Science item-difficulty parameter scatter plot by grade.

Table III-3. Science Item-Parameter Correlations between Female and Male Samples

Grade	Item discrimination	Item difficulty
5	0.96	0.98
8	0.90	0.96
11	0.96	0.92

III.3.3. Differential item functioning (DIF). DIF examines whether an item shows statistical difference between two groups of students after any effect due to ability is removed. Logistic regression was used to detect items with DIF. Based on Jodoin and Gierl's (2001) DIF classification criteria, when the DIF test is significant, moderate DIF has a Nagelkerke R^2 change between 0.035 and 0.070, and large DIF has a Nagelkerke R^2 change greater than 0.070.

DIF was examined across gender (female vs. male) and race (Black vs. White) groups. Tables III-4 through III-6 show the number of items identified as having DIF, by grade, for ELA, mathematics, and science. As seen in the tables, the number of items with DIF is close to or equal to zero for all three subjects.

The low DIF item count is expected because CETE has been proactive in improving item quality. Item statistics are used to help write better items over the years. DIF has been addressed by providing effective item bias and sensitivity training and also guidance to item writers and item reviewers. The concept has been emphasized during item-writing training, item writing, and both

internal and external item reviews. The effort has resulted in a decrease in the numbers of DIF items over time.

Table III-4. ELA DIF Item Count by Grade

Grade	No. of items	Gender DIF		Race DIF	
		Moderate	Large	Moderate	Large
3	86	0	0	0	0
4	82	0	0	0	0
5	81	0	0	0	0
6	85	0	0	1	1
7	74	0	0	0	0
8	80	0	0	0	0
10	68	0	0	0	0

Table III-5. Mathematics DIF Item Count by Grade

Grade	No. of items	Gender DIF		Race DIF	
		Moderate	Large	Moderate	Large
3	88	0	0	0	0
4	80	0	0	0	0
5	87	0	0	0	0
6	85	0	0	0	0
7	80	0	0	0	0
8	83	0	0	0	0
10	83	0	0	0	0

Table III-6. Science DIF Item Count by Grade

Grade	No. of items	Gender DIF		Race DIF	
		Moderate	Large	Moderate	Large
5	42	0	0	0	0
8	48	0	0	0	0
11	47	0	0	0	0

III.4. Validity Based on Relationships to 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" (American Psychological Association et al., 2014, p. 16).

This kind of evidence refers to external evidence and is classified into three types: convergent, discriminant, and criterion related. Convergent evidence is provided by the relationships between students' performance on different assessments intended to measure similar constructs. Discriminant evidence is provided by the 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 on a criterion measure (Cronbach, 1971; Messick, 1989).

III.4.1. Relationships among KAP subjects. Convergence validity requires that another test measures a similar construct; because of the nature of testing for the purposes of state and federal accountability, it is not feasible to administer a test similar to the KAP assessment. Discriminant validity can be evaluated using the correlation between subjects, such as ELA and mathematics. Past studies showed high correlations between subjects, which indicates that some common traits are shared across subjects; however, the correlations should not be too high. The correlations presented in Table III-7 are between subjects of the same grade, and the values range from 0.68 to 0.77. Correlations are not computed between different grades.

Table III-7. Correlations Among ELA, Mathematics, and Science Scores

Grade	ELA vs. mathematics	ELA vs. science	Mathematics vs. science
3	0.77		
4	0.76		
5	0.75	0.75	0.71
6	0.76		
7	0.73		
8	0.73	0.72	0.68
10	0.70		
11			

III.4.2. Relationships between scale scores and demographic variables. Further, discriminant validity was evaluated using the correlations between students' scale scores and their demographic background (i.e., gender, Hispanic, EL, disability, and ethnicity) within subject and grade. As shown in Tables III-8 through III-10, the correlations, except for the one between students' scale scores and their disability status, are very low. Correlations between scale scores and disability group range from −0.58 to −0.40, with most of them around −0.50. The negative relationship indicates that students with disabilities did not perform as well as those without disabilities. Correlations between students' scale scores and their EL status, which range from −0.39 to −0.28, indicate that a lack of English language fluency can affect students' performance in a negative way. The tests are all presented in English; in mathematics and science, additional supports for EL students are provided, such as keyword translation into Spanish or the availability of word-to-word translation dictionaries for EL students. The strength of the relationship between scale scores and Hispanic group identity is slightly weaker but similar to that of the relationship between scale scores and the EL group, with correlations ranging from

−0.31 to −0.26. Non-Hispanic students seem to perform better on ELA, mathematics, and science. There is nearly no relationship between mathematics scores and gender group; however, girls seemed to do slightly better than boys did on ELA (*r* ranging from −0.18 to −0.10), and boys performed slightly better than girls did on grade 8 science (*r* = 0.11).

Table III-8. Correlations Between Scale Scores and Demographic Groups for ELA

Grade	Gender	Hispanic	EL	Disability	Ethnicity
3	−0.10	−0.30	−0.34	−0.43	−0.13
4	−0.11	−0.28	−0.33	−0.47	−0.12
5	−0.13	−0.28	−0.34	−0.51	−0.13
6	−0.11	−0.31	−0.36	−0.55	−0.14
7	−0.13	−0.28	−0.35	−0.55	−0.15
8	−0.16	−0.27	−0.35	−0.58	−0.13
10	−0.18	−0.29	−0.39	−0.55	−0.17

Table III-9. Correlations Between Scale Scores and Demographic Groups for Mathematics

Grade	Gender	Hispanic	EL	Disability	Ethnicity
3	0.04	−0.28	−0.29	−0.41	−0.09
4	0.07	−0.28	−0.29	−0.42	−0.10
5	0.05	−0.28	−0.29	−0.45	−0.08
6	0.03	−0.29	−0.29	−0.46	−0.09
7	0.03	−0.28	−0.30	−0.52	−0.10
8	0.00	−0.26	−0.28	−0.50	−0.07
10	0.00	−0.28	−0.30	−0.45	−0.09

Table III-10. Correlations Between Scale Scores and Demographic Groups for Science

Grade	Gender	Hispanic	EL	Disability	Ethnicity
5	0.06	−0.28	−0.33	−0.40	−0.13
8	0.11	−0.30	−0.36	−0.44	−0.18
11	0.08	−0.29	−0.39	−0.43	−0.17

III.4.3. Relationships between KAP scores and ACT scores. A predictive study between the KAP and ACT scores was conducted in fall 2016. According to ACT, the ACT test measures what students learn in high school, and ACT scores are used to determine students' academic readiness for college. KAP adopted the KCCRS, which are also an indicator of college readiness. Scores of the two tests refer to somewhat different content specifications but have the same intention. Among the ACT scores, English, reading, mathematics, and composite scores (the average of scores of the four multiple-choice subjects: English, mathematics, reading, and science) were used to correlate with the KAP ELA and mathematics scores.

This study used student ACT scores from 10 school districts. After data cleaning, about 5,369 ACT scores taken after the KAP spring 2015 administration were kept to analyze with 2015 KAP scores. When a student had multiple ACT scores, only one score was selected. Two score selection approaches were used: the first composite score and the highest composite score. The

first ACT score was used because its testing date was closer to the KAP testing window. The highest score was used because it is typically accepted by colleges regardless of the number of times students took the test. Results produced from these two samples (i.e., the first ACT score sample and the highest ACT score sample) are reported in Table III-11.

As shown in Table III-11, the correlation between the tests is greater than 0.62; the highest correlation of ACT and KAP scores is 0.85. Logically, KAP ELA scores correlate better with ACT English and reading scores than with ACT mathematics scores, and KAP mathematics correlates better with ACT math scores. Both KAP ELA and mathematics scores correlate well with ACT composite scores (0.77–0.79).

*Table III-11. Correlations Among KAP and ACT Scores (*N = 5,369)

	KAP correlation with first ACT scores		KAP correlation with highest ACT scores	
	ELA	Mathematics	ELA	Mathematics
ACT score				
Composite	0.78	0.78	0.77	0.79
English	0.77	0.69	0.76	0.70
Reading	0.73	0.61	0.73	0.62
Math	0.64	0.85	0.64	0.85

Figures III-3 and III-4 are comparison graphs between the percentage of students whose scores were Level 3 or Level 4 (Level 3/4) on the KAP assessment for grade 10 ELA and mathematics and the percentage of students meeting ACT benchmarks from 2015 to 2017. As shown in the graphs below, the KAP Level 3/4 percentages for grade 10 ELA are 32%, 32%, and 30% for 2015, 2016, and 2017, respectively, while the ACT's benchmark meeting rates for English are 71%, 70%, and 69%, respectively. The KAP Level 3/4 percentages for grade 10 mathematics are 25%, 24%, and 25% for 2015, 2016, and 2017, respectively, while the ACT's benchmark meeting rates for mathematics are 49%, 48%, and 46%, respectively.

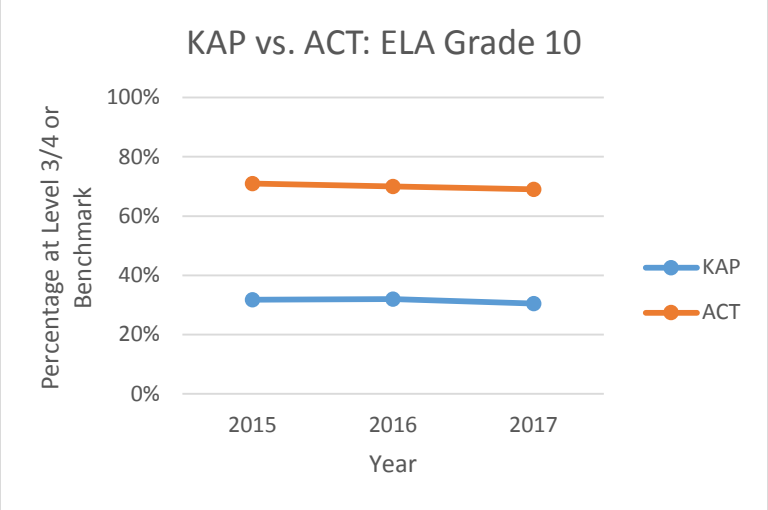


Figure III-3. Grade 10 ELA trends across years: KAP vs. ACT.

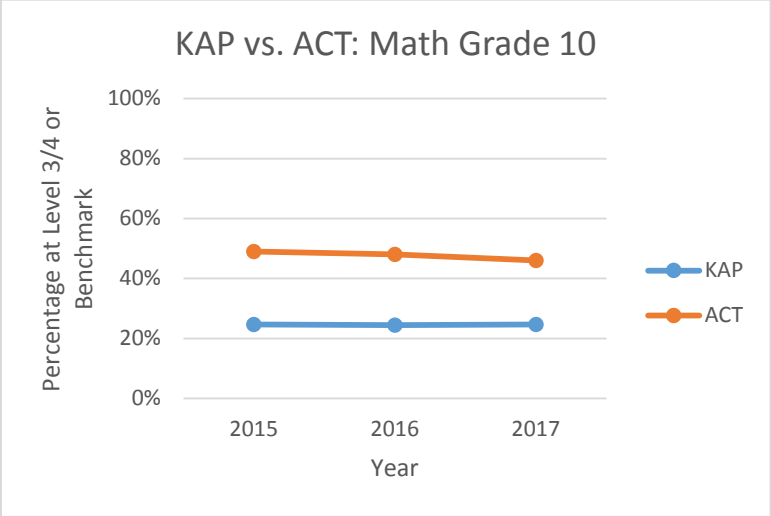


Figure III-4. Grade 10 mathematics trends across years: KAP vs. ACT.

III.4.4. Relationships between the KAP assessment and the National Assessment of Educational Progress (NAEP). The state of Kansas participates in NAEP, otherwise known as the Nation's Report Card. NAEP is considered the gold standard of assessments. It is the largest nationally representative assessment of what American students know and can do, and it serves a different role than state assessments. The NAEP assessments allow each state to be compared to national results and to evaluate progress over time. It informs the public about the academic achievement of elementary and secondary students in Kansas and in the United States. For more details, visit the KSDE website at http://www.ksde.org/Agency/Division-of-Learning-Services/Career-Standards-and-Assessment-Services/CSAS-Home/Assessments/National-Assessment-of-Educational-Progress-NAEP.

Comparisons between KAP Level 3/4 rates and the corresponding rates of Proficient/Advanced (P/A) on NAEP across years for grades 4 and 8 ELA and mathematics are presented in Figures III-5 through III-8. In years 2015 through 2017, KAP Level 3/4 rates ranged from 50% to 56% for grade 4 ELA, from 28% to 32% for grade 8 ELA, from 36% to 40% for grade 4 mathematics, and from 24% to 27% for grade 8 mathematics. The Level 3/4 rates of the Kansas NAEP and the P/A rates on the national NAEP for both grade 4 and grade 8 ELA are very similar across years (i.e., the odd-numbered years from 2003 to 2015), ranging from 30% to 40%, with most of them around 35%. However, the Level 3/4 rate of Kansas NAEP is consistently higher than that of the national NAEP P/A rate over the years (i.e., the odd-numbered years from 2003 to 2015) for both grade 4 and grade 8 mathematics.

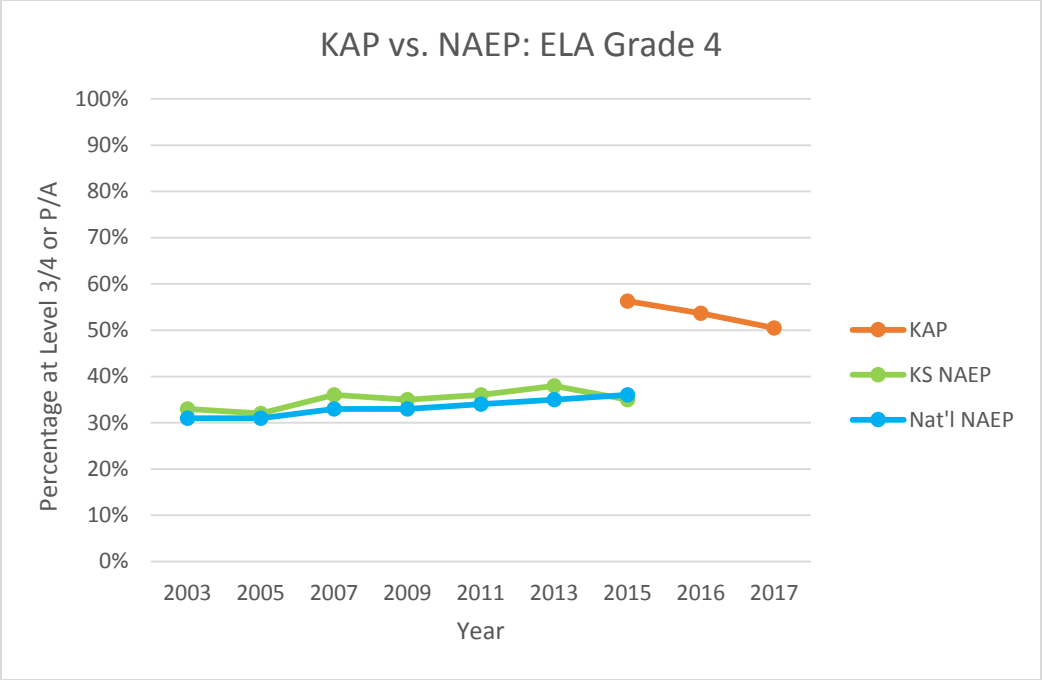


Figure III-5. Grade 4 ELA trend across years: KAP vs. NAEP.

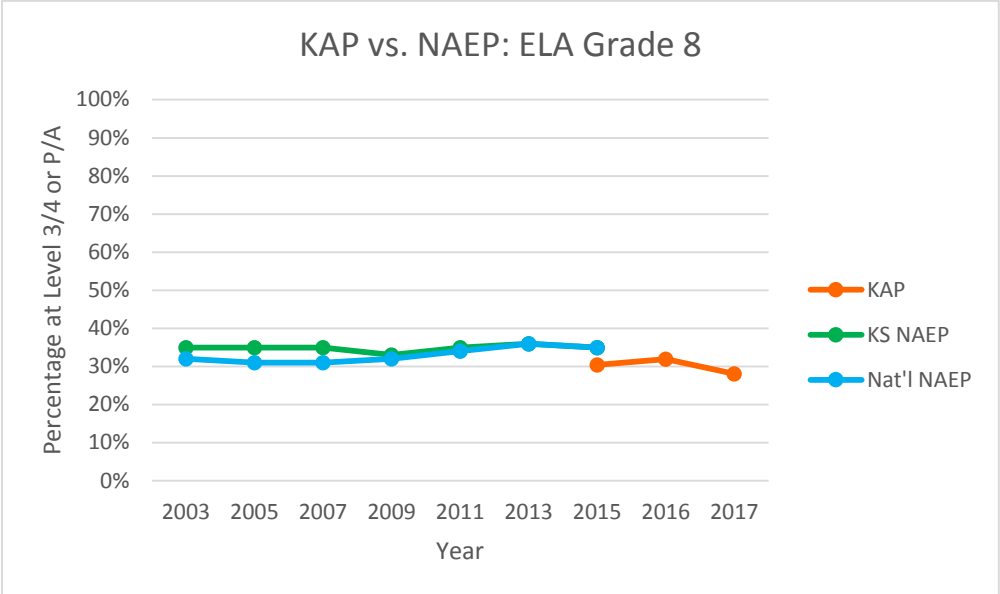


Figure III-6. Grade 8 ELA trend across years: KAP vs. NAEP.

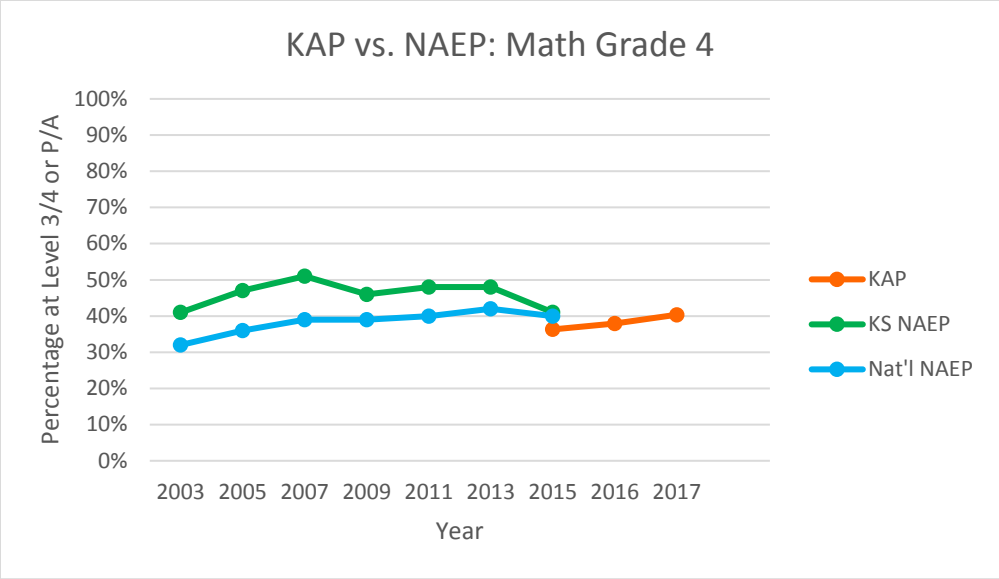


Figure III-7. Grade 4 mathematics trend across years: KAP vs. NAEP.

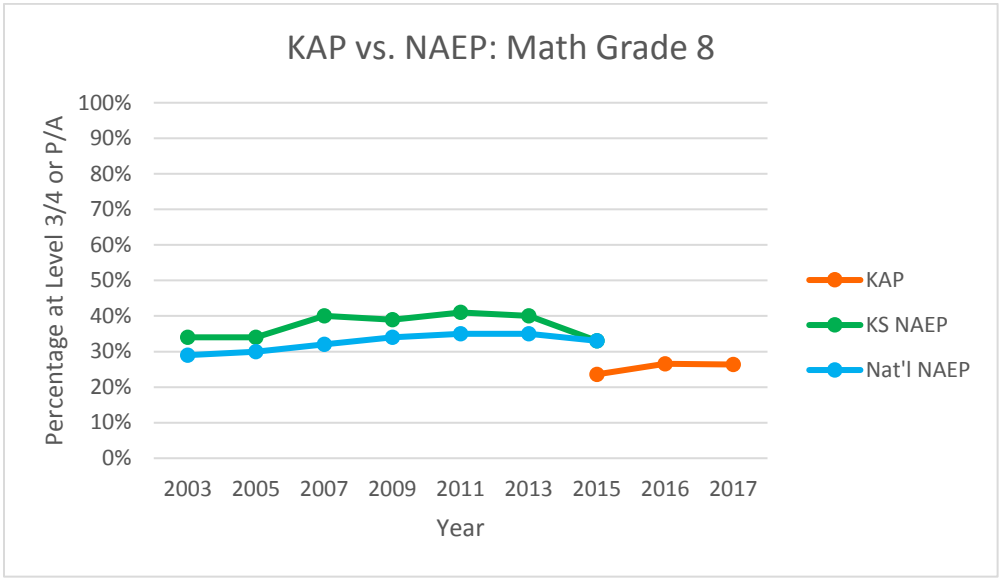


Figure III-8. Grade 8 mathematics trend across years: KAP vs. NAEP.

III.5. Survey Results for Validity Evidence

KSDE is providing perception surveys, free of charge, to schools/districts interested in gathering perception data from students, staff, and community members. The purpose of conducting this survey is to gain feedback from Kansas educators to improve the KAP assessment in ELA, mathematics, science, HGSS, and the Kansas English Language Proficiency Assessment (KELPA). The survey was administered during April and May 2017 to Kansas educators in

Qualtrics survey software. Survey questionnaires were sent to Kansas educators via the Assessment list serve and the Curriculum Leaders list serve. District and Building Test Coordinators, District and Building Administrators, and teachers were surveyed. The survey results are presented in Appendix G.

In addition to rating and multiple-choice survey questions, one open-ended question was asked to gain educators' comments on the following question. If you could change one thing about the assessment, what change would you make? Other than some general comments, participants' feedback fell into the following categories:

- Changes for the length of the test (i.e., to make it shorter);
- Changes in results presentation, e.g.,
    - Allow educators to see results sooner;
    - Provide the results to the students instantly;
    - Get faster feedback on results;
    - Would like to see immediate results as before;
    - Quicker turn around for test results;
- Changes needed for students with disabilities;
- Changes regarding technology; and
- Changes regarding the website.

IV. Technical Quality—Others

IV.1. Reliability

Reliability is a test score consistency index. It is based on the sampling theory that a test is only a sample of all possible items in a content area. To use test scores to infer the knowledge and skills of the content area, the tested content must be representative of the entire content area as defined by the content standards. Additionally, factors that can affect performance—such as allocated testing time, computer environment, and supporting materials—should be standardized to remove undesirable effects. The *Standards for Educational and Psychological Testing* (American Psychological Association et al., 2014) states that the first step in examining test reliability is to investigate the specifications of replications of the testing procedures. KAP has standardized its testing procedures, and the same procedures are applied to all students; specific accommodations are provided to students with special needs. The testing specifications can be found in the *Examiner's Manual*.

Because reliability theory defines each test form as only a sample of the tested content area, different test forms of a subject are different samples of the content area and may yield different observed scores. In sampling theory, the mean of repeated samples' means can be used to infer the population mean. In testing theory, the mean of repeated testing scores is the test taker's true score of the defined content area. However, it is impractical to test the same content area repeatedly because test takers cannot maintain the same knowledge, physical condition, and mental status across test administrations. Factors such as learning, fatigue, and motivation may affect test takers at different rates, making an empirical study of reliability unlikely in the context of an educational achievement test. Therefore, a reliability index that is derived through theories must similarly be approached from a more theoretical standpoint.

A fundamental reliability theory is defined by the classical test theory. Classical test theory has established that any observed score is the composite of the true score and some amount of measurement error. Measurement error can be caused by factors, such as a differential effect attributable to learning or differences in motivation, among individuals. Typically, reliability values range from 0 to 1. Higher values indicate better test reliability.

IV.1.1. Test reliability. ELA and mathematics tests use IRT models to estimate students' latent proficiency (theta), which is then transferred to a scaled score. A standard error (*SE*) is also estimated for each value of theta and is then transformed to the conditional standard error of measurement (CSEM). CSEMs are computed through their inverse relationship with test information functions. Graphical representations of CSEM curves can be found in Appendix B. The information function and the CSEM are computed using all operational items in a grade, not by block or path. Typical CSEM values are low at the center of the scale-score distribution and gradually increase toward the two ends of the scale, whereas scaled scores become very low or very high and result in a U-shaped pattern. However, some CSEM curves presented in Appendix B have lower values at the low scaled-score side, which reflects improved measurement precision at the lower end of the distribution because of the stage-adaptive model and inclusion

of a sufficient number of items with lower difficulty levels. Tables C-1 to C-17 in Appendix C present scale scores with associated CSEMs and their frequencies for each subject and grade. Test reliability by grade and subject is presented in Table IV-1.

The *SE*s and their scaled values, CSEMs, indicate reliability by scaled-score points. Green, Bock, Humphreys, Linn, and Reckase (1984) used the *SE*s of theta (θ) to derive an index for test-level reliability:

$$\bar{\rho} = \frac{\sigma_\theta^2 - \overline{SE_\theta^2}}{\sigma_\theta^2}. \qquad \text{(IV-1)}$$

Green et al. (1984) called this index marginal reliability. The equation shows that marginal reliability, $\bar{\rho}$, is defined by two values: the variance of theta (σ_θ^2) and *SE*s of theta (SE_θ^2). Because *SE*s are different across thetas, the mean of squared *SE*s, $\overline{SE_\theta^2}$, is used in the equation.

Table IV-1. Test Reliability by Grade and Subject

	Subject		
Grade	ELA	Mathematics	Science
3	0.92	0.94	
4	0.90	0.94	
5	0.91	0.94	0.82
6	0.91	0.94	
7	0.90	0.92	
8	0.90	0.93	0.83
High School	0.92	0.93	0.86

Reliabilities of ELA and mathematics tests are above 0.90. This high reliability range may reflect the benefit of multistage design. The science test has relatively low reliability because of fewer test items compared to ELA and mathematics but is still above 0.80.

IV.1.2. Classification consistency and accuracy. How accurately students are classified into performance categories is of a great interest for accountability testing programs. Classification consistency refers to the agreement between two parallel forms, and classification accuracy refers to the agreement between true scores and observed scores (Livingston and Lewis, 1995). Tables IV-2 and IV-3 present the possible classification results of consistency and accuracy of two performance levels, respectively. Both tables indicate that when students are classified into two levels, four possible outcomes are yielded, respectively, by the parallel forms and by the true scores and observed scores. Among the four possible outcomes, two of them are consistent (accurate), and two of them are inconsistent (false).

Table IV-2. Cross-Tabulation of Classification Consistency

		Observed score Parallel form 2	
		Level X	Level Y
Observed score Parallel form 1	Level X	Consistent classification	Inconsistent classification
	Level Y	Inconsistent classification	Consistent classification

Table IV-3. Cross-Tabulation of Classification Accuracy

		True score	
		Level X	Level Y
Observed score	Level X	Accurate classification	False negative
	Level Y	False positive	Accurate classification

As mentioned previously, true scores are unobservable, and repeat testing is not feasible. In order to evaluate the classification consistency and accuracy of single administration, alternative statistical procedures have been developed. Among them, Livingston and Lewis (1995) described procedures that are broadly used because they are not limited to dichotomous items and do not assume equal weight on items. The Livingston and Lewis method uses (a) test reliability to estimate "effective length," (b) a user-selected true score model to predict the parallel form's observed score distribution for consistency comparison, and (c) a user-selected model to predict the true score distribution for accuracy estimates.

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 Table IV-4. BB-CLASS software (Brennan, 2004) was used to derive the information. All science tests are shorter; therefore, they have comparatively lower classification outcomes.

Table IV-4. Classification Consistency and Accuracy by Subject and Grade

	Cut-score category							
	Overall		1 vs. 2, 3, 4		1, 2 vs. 3, 4		1, 2, 3 vs. 4	
Grade	Consistency	Accuracy	Consistency	Accuracy	Consistency	Accuracy	Consistency	Accuracy
ELA								
3	0.60	0.79	0.72	0.92	0.76	0.92	0.74	0.95
4	0.57	0.79	0.60	0.94	0.73	0.90	0.70	0.95
5	0.57	0.77	0.69	0.92	0.74	0.91	0.72	0.94
6	0.59	0.79	0.73	0.92	0.74	0.91	0.62	0.96
7	0.57	0.78	0.70	0.91	0.74	0.91	0.68	0.96
8	0.59	0.80	0.69	0.92	0.72	0.92	0.62	0.97
10	0.62	0.81	0.73	0.92	0.75	0.92	0.66	0.97
Mathematics								
3	0.64	0.81	0.66	0.94	0.79	0.93	0.79	0.96
4	0.64	0.83	0.60	0.93	0.81	0.93	0.79	0.97
5	0.65	0.82	0.71	0.92	0.81	0.94	0.79	0.97
6	0.63	0.82	0.66	0.91	0.82	0.94	0.80	0.97
7	0.60	0.81	0.59	0.90	0.78	0.93	0.72	0.98
8	0.64	0.83	0.74	0.91	0.80	0.95	0.76	0.98
10	0.63	0.82	0.72	0.91	0.81	0.95	0.79	0.98
Science								
5	0.40	0.67	0.52	0.88	0.64	0.87	0.62	0.93
8	0.44	0.70	0.60	0.87	0.65	0.89	0.60	0.94
11	0.48	0.72	0.61	0.88	0.70	0.90	0.68	0.95

IV.1.3. Subgroup reliability. Subgroup reliabilities are presented in Appendix D. Marginal reliability is used. Appendix D shows that the race analysis has a smaller sample size than other subgroups because students whose demographic information about race was not provided were excluded from the analysis.

Both ELA and mathematics have very high subgroup reliabilities. The ELA subgroup reliabilities are around 0.90, with the majority of them in the lower 0.90 range and a few in the upper 0.80 range. Mathematics subgroup reliabilities are in the mid-0.90 range. Science has lower subgroup reliabilities because of fewer items, but they are close to overall science test reliabilities. They range from 0.78 to 0.88, with most of them in the mid-0.80 range.

IV.1.4. Path reliability. Path reliability is the product of using a multistage adaptive test design, thus, it applies only to ELA and mathematics tests. The multistage adaptive test design dictates that different sets of items (blocks) are assigned to students at Stage 2. The different paths mean that students take item sets with different levels of difficulty. Analytically, multiple test forms are taken by students. Conceptually, path reliability is equivalent to the reliability of different test forms. The results of path reliability can be found in Appendix E.

The stages in Appendix D tables provide block information for each stage, student count and percentage, and reliability. For example, the path reliability of the ELA test in grade 3, presented in Table IV-5, indicates two paths (forms). Stage 1 has only one block of items, with a medium level of difficulty. Stage 2 has two blocks of items: easy and hard.

Table IV-5. ELA Grade 3 Path Reliability

Path	Stage 1	Stage 2	*N*	Percentage	Reliability
			38,314		
1	Medium	Easy	20,360	53.1%	0.93
2	Medium	Hard	17,954	46.9%	0.92

IV.1.2. Subscore reliability. Besides the total test score, scores of subsets of ELA, mathematics, and science items are also reported for students. The number of items in each subscore category varies; additionally, some items contribute to multiple subscores. The minimum number of items reported for a subscore is six.

ELA has a total of nine subscores, and all of the ELA grades report the same nine subscores. The primary subscores are Claim 1 (Reading) and Claim 2 (Writing). A second set of subscores within Claim 1 can be used only to compare outcomes when the text that is used as the stimulus is from a literary genre (such as narrative or poetry) or is an informational text. A third set of subscores, also within Claim 1, looks at performance on targets or combinations of targets (across both literary and informational texts) that measure "main ideas and supporting details" and "making and supporting inferences and conclusions." A final set of subscores, from within Claim 2, compare performance on targets or combinations of targets that measure "revising texts," "language and vocabulary use," and "grammar and conventions."

The number of mathematics subscores varies across grades. Grade 3 has six subscores; grade 4 has eight subscores; grade 5 has seven subscores; grade 6 has six subscores; grade 7 has seven subscores; and grades 8 and 10 have six subscores. All grades include four separate subscores for Claims 1, 2, 3, and 4. The additional subscores that can be reported are shown in Table IV-6.

Table IV-6. Additional Subscores for ELA by Grade

	3	4	5	6	7	8	10
1–Overall Concepts and Procedures	X	X	X	X	X	X	X
1.1–Operations and Algebraic Thinking	X	X					
1.2–Number and Operations in Base Ten		X	X				
1.3–Number and Operations with Fractions		X	X				
1.4–Measurement and Data	X	X	X				
1.5–The Number System				X			
1.6–Expressions and Equations				X	X	X	
1.7–Algebra							X
1.8–Geometry					X	X	X
1.9–Statistics and Probability					X		X
2–Problem Solving	X	X	X	X	X	X	X
3–Communicating Reasoning	X	X	X	X	X	X	X
4–Modeling and Data Analysis	X	X	X	X	X	X	X

Science has three subscores corresponding to Claims 1, 2, and 3. These are Physical Science, Life Science, and Earth and Space Science. Raw score ranges for each claim and grade are different. Because of the shorter length of the test, additional subscores at a finer grain size, such as those for ELA and mathematics, are not reported for science.

These subscores are reported in three categories: below, meets, and exceeds. When a student respond to less than 60% of the items in a claim, *insufficient data* will be reported instead of a subscore category. Subscore categories are assigned according to scaled scores calculated on the subscores. The procedure for computing subscore scaled scores is similar to that for computing test scaled scores: Student latent proficiencies (thetas) in each subscore category are estimated using IRT models and are then linearly transformed to scaled scores using the test's scaling constants. Item parameters derived at the test level are used to derive thetas for subscores. Cuts of 300 and 325 (one *SE* above 300) are chosen to define students' subscore categories. Subscore scale scores less than 300 are "below," 300 to 325 are "meets," and above 325 are "exceeds."

Two analyses are conducted to determine the reliability of subscores. First, the subscore marginal reliabilities are computed. In general, statistical estimations are affected by sample sizes. For a test, estimations are affected by both the number of items and student sample sizes. Because the KAP assessment is given to a large number of students, the subscore reliability is mainly driven by number of items. It is expected that the reliability of some subscores may be affected by smaller item counts. Second, the classification consistency and accuracy of subscores are examined because the subscores are reported in categories.

Table IV-7 reports a summary of the subscore reliability and classification consistency and accuracy. Most subscore reliabilities are within good range. The average consistency indices range from 0.31 to 0.35, and the average of accuracy indices is around 0.70 for all three subjects.

Table IV-7. Summary of Subscore Reliability and Classification Consistency and Accuracy by Subject

Subject	No. of subscores	*M*	*SD*	Min	P_{25}	P_{50}	P_{75}	Max
Reliability								
ELA	63	0.62	0.06	0.50	0.58	0.62	0.67	0.74
Mathematics	46	0.62	0.08	0.48	0.56	0.61	0.67	0.77
Science	9	0.58	0.02	0.56	0.57	0.59	0.59	0.61
Consistency								
ELA	63	0.34	0.05	0.23	0.31	0.34	0.38	0.46
Mathematics	46	0.35	0.08	0.21	0.29	0.34	0.40	0.53
Science	9	0.31	0.03	0.26	0.29	0.3	0.32	0.35
Accuracy								
ELA	63	0.71	0.05	0.60	0.67	0.70	0.75	0.81
Mathematics	46	0.72	0.08	0.43	0.69	0.72	0.76	0.86
Science	9	0.69	0.04	0.62	0.67	0.70	0.70	0.74

Note. P_{25} = 25th percentile; P_{50} = 50th percentile; P_{75} = 75th percentile.

IV.2. 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" (American Psychological Association et al., 2014, p. 74). This 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, such as item and test development, inclusion and accommodations, and DIF.

UD was used as a guide during the development of items, test formats, and the online test delivery interface. UD refers to principles that provide equal access to all students. While initially designed to meet the interests of students with special needs, universally designed assessments provide benefits to all students. Implementation of UD started during item-writer training. Using appropriate item- and test-development processes is an excellent start to help ensure fairness. However, some barriers, such as blindness, cannot be addressed by UD. Test inclusion and accommodations policies help address these needs. Many accommodations are provided in the online test system, including magnification, text-to-speech, and image contrasts, among others. Some students will require braille tests, which are made available to students who need them. (For details about accommodations, see V: Inclusion of All Students.)

Further evidence of the fairness and accessibility of the KAP assessment is seen in DIF analysis. DIF analysis examines whether an item shows any statistical difference between two groups of students after controlling for student proficiency. The DIF analysis results presented in III: Technical Quality—Validity show that, out of nearly 2,000 operational items for all subjects and grades, only one item shows moderate DIF and one shows large DIF.

IV.3. Full Performance Continuum

The KAP assessment was developed with the goal that assessment of each subject area and grade level would provide a reasonably precise estimation of student proficiency across the full performance continuum (i.e., from low-performing to high-performing students). This goal is fulfilled by using items that cover different DOK levels and a wide range of difficulties. As mentioned earlier, although the proportions of each DOK level are not specified in the test blueprints, the expected DOK level is explicitly stated in the item specifications. When test items are written to each assessment target, the items also have to reflect the expected DOK level as implied by the content to be measured. This expectation is emphasized throughout the item writing and during both internal and external item reviews. Consequently, when the items selected for a test meet the blueprint, those items also meet the underlying DOK requirements.

During test construction, there is no constraint on item *p* values or mean scores. Item quality is screened through item-total correlation, DIF, option analyses, and IRT parameters. This approach not only ensures the quality of items to be used on the test but also provides the widest range possible in measuring student abilities. Additionally, curves of test characteristic, test information, and CSEM are plotted during test construction to gauge the proficiency range each test covers. Note that one of the advantages of the adaptive test design is that it enables the test to extend from the extremely low- to the high-ability range that is typically ruled out by a non-adaptive design. To confirm that the tests efficiently cover the full performance continuum as expected, classical and IRT item statistics are presented here as evidence.

IV.3.1. Classical item statistics. Two statistics, item difficulty and item discrimination, are calculated and provided. Item difficulty refers to how easy or difficult an item is, and item discrimination indicates the degree to which an item differentiates between students with high proficiency and those with low proficiency. Item difficulty in classical test theory is expressed as a *p* value or mean score. A *p* value is the percentage of students who answered the item correctly. Equation IV-1 shows the calculation of the *p* value.

$$p \text{ value / average proportion score} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i}{\text{item max score}}, \qquad \text{(IV-1)}$$

where *x* refers to the observed score, *i* refers to student *i*, and *n* refers to the total number of students who took the item.

For difficult multiple-choice items with four response options, complete random guessing by students would lead to an expected *p* value of ¼ point (0.25). That suggests that there is a 25% chance that a student will guess the correct response without any related prior knowledge. For multiple-choice items with five response options, the guessing *p* value would be ⅕ point (0.20), and so on, for other numbers of response options. However, the thoughtful development of incorrect answer choices can lead to much lower than theoretical asymptotes. This strategy also leads to poor model fit of the three-parameter logistic (3PL) IRT model, which assumes monotonically increasing item characteristic curves (ICCs), and these very attractive distractors

can result in nonmonotonic ICCs. Thus, CETE typically uses the 2PL and its polytomous counterpart, the GRM.

Summaries of item difficulty for ELA, mathematics, and science tests are presented in Tables IV-8 through IV-10. The ELA grade-level average item difficulties range from 0.51 to 0.55; the mathematics grade-level average item difficulties range from 0.46 to 0.55; and the science grade-level average item difficulties range from 0.52 to 0.58. Note that P_{25} and P_{75} in the following tables refer to the 25th and 75th percentiles, respectively.

Table IV-8. Summary Statistics for Classical Item Difficulties for ELA

Grade	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
3	86	0.51	0.16	0.15	0.40	0.53	0.62	0.86
4	82	0.52	0.17	0.15	0.41	0.51	0.63	0.93
5	81	0.53	0.13	0.15	0.45	0.53	0.64	0.78
6	85	0.54	0.17	0.16	0.41	0.55	0.66	0.88
7	74	0.55	0.17	0.23	0.44	0.52	0.66	0.95
8	80	0.55	0.17	0.10	0.48	0.57	0.67	0.82
10	68	0.54	0.17	0.02	0.44	0.55	0.65	0.86

Table IV-9. Summary Statistics for Classical Item Difficulties for Mathematics

Grade	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
3	88	0.54	0.20	0.09	0.39	0.52	0.68	0.92
4	80	0.55	0.17	0.08	0.43	0.58	0.66	0.87
5	87	0.48	0.17	0.09	0.37	0.50	0.61	0.81
6	85	0.47	0.16	0.07	0.39	0.49	0.57	0.84
7	80	0.48	0.14	0.10	0.38	0.47	0.58	0.78
8	83	0.46	0.18	0.02	0.36	0.48	0.59	0.89
10	83	0.47	0.18	0.04	0.35	0.48	0.57	0.95

Table IV-10. Summary Statistics for Classical Item Difficulties for Science

Grade	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
5	42	0.58	0.17	0.25	0.48	0.57	0.72	0.90
8	48	0.53	0.12	0.26	0.47	0.53	0.61	0.77
11	47	0.52	0.14	0.18	0.46	0.54	0.60	0.79

Item discrimination reflects an item's ability to differentiate students of high proficiency from those of low proficiency. Ideally, high-achieving students (i.e., those with high raw scores) should be more likely to answer any given item correctly, whereas low-achieving students (i.e., those with low raw scores) should be more likely to answer the same item incorrectly. The Pearson's product-moment correlation coefficient between student item scores and test scores is also referred to as item-total correlations, although strictly speaking these are point-biserial

correlations when items have dichotomous (0, 1) scores. The item-total correlation is used as an index of item discrimination. The item-total correlation ranges from −1.0 to 1.0. Positive values indicate that students with higher raw scores are more likely to answer an item correctly than those with low rawer scores; negative values indicate the opposite. The magnitude of the correlation indicates the degree of discrimination in that items with higher values have better discrimination power.

Tables IV-11 through IV-13 present item discrimination for the three subjects. The medians of item discrimination for ELA range from 0.32 to 0.37; the medians of item discrimination for mathematics range from 0.36 to 0.39; and the medians of item discrimination for science range from 0.35 to 0.39.

Table IV-11. Summary Statistics for Classical Item Discrimination for ELA

Grade	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
3	86	0.37	0.08	0.11	0.31	0.37	0.42	0.56
4	82	0.35	0.09	0.10	0.31	0.36	0.41	0.59
5	81	0.34	0.09	0.07	0.29	0.35	0.40	0.54
6	85	0.36	0.09	0.15	0.29	0.37	0.42	0.55
7	74	0.32	0.08	0.14	0.26	0.32	0.38	0.54
8	80	0.33	0.11	−0.05	0.24	0.34	0.41	0.53
10	68	0.36	0.11	0.13	0.30	0.35	0.44	0.64

Note. P_{25} = 25th percentile; P_{75} = 75th percentile.

The grade 8 item with a negative biserial (it is a multiple-choice item) in 2017 had a biserial of 0.228 during its calibration year (2016). All statistics used in test routing and scoring were based on the pre-equated statistics from the calibration year. While the difference in value from the calibration year to the current test administration is very unexpected, content experts concluded that there was no error in scoring that crept in, and the item remained unchanged from year to year. Psychometricians and content experts will continue to monitor this item if it is used again.

Table IV-12. Summary Statistics for Classical Item Discrimination for Mathematics

Grade	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
3	88	0.39	0.10	0.16	0.32	0.38	0.46	0.57
4	80	0.40	0.09	0.22	0.34	0.39	0.48	0.61
5	87	0.39	0.09	0.13	0.33	0.39	0.45	0.58
6	85	0.39	0.09	0.18	0.32	0.38	0.45	0.56
7	80	0.36	0.10	0.13	0.30	0.36	0.43	0.55
8	83	0.37	0.09	0.09	0.32	0.38	0.43	0.56
10	83	0.36	0.10	0.06	0.29	0.37	0.43	0.56

Note. P_{25} = 25th percentile; P_{75} = 75th percentile.

Table IV-13. Summary Statistics for Classical Item Discrimination for Science

Grade	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
5	42	0.37	0.08	0.19	0.33	0.38	0.43	0.52
8	48	0.34	0.07	0.17	0.30	0.35	0.40	0.46
11	47	0.37	0.10	0.13	0.31	0.39	0.43	0.54

Note. P_{25} = 25th percentile; P_{75} = 75th percentile.

IV.3.2. IRT item statistics. Tables IV-14 through IV-19 summarize the difficulty (i.e., *b* parameter) and discrimination (i.e., *a* parameter) estimates of operational items in ELA, mathematics, and science tests, respectively. Most items are dichotomous, but some items have as many as 11 score categories (thus, 10 *b* parameters yet still only one *a* parameter); therefore, the numbers of *b* and *a* parameters are different in these tables. Parameters for all items, irrespective of the number of score categories, are included together in the tables below.

The mean item difficulty increases as the grade increases for science. The mean item difficulty remains similar from grade 3 to grade 8, but it increases dramatically in grade 10 for mathematics. The mean item difficulty fluctuates across grades. A large standard deviation (*SD*) of difficulty parameters indicates a large variability of item difficulties. The minima and maxima for the difficulty parameters indicate that the items included in KAP assessments adequately cover the full performance continuum. Although item discrimination is not usually too far from 1.0 on average, it clearly varies over items, justifying the use of the 2PL that permits the discrimination parameter to vary over items. The median item discrimination declines as the grade increases for mathematics; however, for ELA and science, it fluctuates a bit but still shows the trend of decreasing as the grade increases. Overall, mathematics has better discrimination parameters than ELA and science.

Table IV-14. Summary Statistics for IRT Item Difficulty for ELA

Grade	No. of *b* parameters	*M*	*SD*	Min	Q1	Median	Q3	Max
3	101	−0.31	1.29	−4.80	−1.08	−0.35	0.41	2.50
4	95	−0.62	1.27	−3.87	−1.55	−0.70	0.14	3.15
5	94	−0.33	1.29	−4.69	−0.98	−0.50	0.48	2.58
6	105	−0.85	1.55	−4.27	−1.89	−0.92	0.13	3.48
7	97	−0.21	2.03	−5.03	−1.26	−0.64	0.82	7.94
8	101	−0.48	1.47	−5.55	−1.18	−0.52	0.35	2.92
10	89	−0.35	1.54	−3.86	−1.34	−0.50	0.53	4.26

Note. *b* = difficulty parameter; Q1 = first quartile; Q3 = third quartile.

Table IV-15. Summary Statistics for IRT Item Difficulty for Mathematics

Grade	No. of *b* parameters	*M*	*SD*	Min	Q1	Median	Q3	Max
3	101	−0.02	1.36	−3.21	−0.96	0.02	0.74	3.38
4	114	−0.10	1.39	−3.45	−1.00	−0.10	0.81	4.08
5	109	−0.01	1.32	−3.65	−0.65	0.04	0.96	3.38
6	96	0.09	1.45	−6.60	−0.54	0.06	0.96	6.02
7	107	0.00	1.73	−4.98	−0.78	0.19	1.11	4.55
8	115	−0.02	2.02	−6.28	−0.90	0.10	1.31	5.00
10	107	0.38	1.73	−4.29	−0.52	0.30	1.35	5.54

Note. *b* = difficulty parameter; Q1 = first quartile; Q3 = third quartile.

Table IV-16. Summary Statistics for IRT Item Difficulty for Science

Grade	No. of *b* parameters	*M*	*SD*	Min	Q1	Median	Q3	Max
5	47	−0.71	1.53	−5.10	−1.41	−0.63	−0.08	2.91
8	59	−0.54	1.82	−6.09	−1.04	−0.20	0.22	3.50
11	54	−0.06	1.66	−5.75	−0.69	−0.27	0.57	5.93

Note. *b* = difficulty parameter; Q1 = first quartile; Q3 = third quartile.

Table IV-17. Summary Statistics for IRT Item Discrimination for ELA

Grade	No. of *a* parameters	*M*	*SD*	Min	Q1	Median	Q3	Max
3	86	1.04	0.36	0.26	0.77	0.99	1.3	1.92
4	82	0.98	0.34	0.34	0.76	0.91	1.22	1.99
5	81	0.91	0.32	0.40	0.71	0.90	1.06	2.43
6	85	1.00	0.34	0.36	0.78	0.96	1.14	2.01
7	74	0.89	0.41	0.26	0.59	0.80	1.18	2.11
8	80	0.91	0.35	0.29	0.67	0.91	1.08	2.00
10	68	0.94	0.32	0.35	0.72	0.94	1.12	2.21

Note. *a* = discrimination parameter; Q1 = first quartile; Q3 = third quartile.

Table IV-18. Summary Statistics for IRT Item Discrimination for Mathematics

Grade	No. of *a* parameters	*M*	*SD*	Min	Q1	Median	Q3	Max
3	88	1.17	0.36	0.51	0.92	1.14	1.4	2.11
4	80	1.15	0.35	0.64	0.85	1.09	1.34	2.02
5	87	1.20	0.31	0.38	1.01	1.17	1.41	1.97
6	85	1.17	0.43	0.43	0.86	1.11	1.44	2.36
7	80	0.99	0.34	0.32	0.76	0.94	1.21	1.93
8	83	1.02	0.36	0.41	0.75	0.96	1.19	2.04
10	83	1.01	0.38	0.33	0.77	0.91	1.26	2.17

Note. *a* = discrimination parameter; Q1 = first quartile; Q3 = third quartile.

Table IV-19. Summary Statistics for IRT Item Discrimination for Science

Grade	No. of *a* parameters	*M*	*SD*	Min	Q1	Median	Q3	Max
5	42	0.84	0.31	0.27	0.65	0.81	1.03	1.51
8	48	0.70	0.21	0.19	0.54	0.71	0.87	1.17
11	47	0.86	0.38	0.26	0.63	0.86	0.97	1.82

Note. *a* = discrimination parameter; Q1 = first quartile; Q3 = third quartile.

IV.3.3. Cognitive complexity. KAP assessment items are categorized by cognitive complexity, as described by Webb's DOK model (Webb, 1997). A description of Webb's DOK follows.

- 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.

Item cognitive complexity is affected by the familiarity of the constructs being measured. Constructs taught previously in the same grade or earlier than described by the KCCRS are likely to appear easier in the early years of the assessment than constructs taught previously in higher grades or not addressed in previous content standards. The DOK associated with each content standard identifies the maximum DOK for an item. Items at Level 4, extended thinking, are not typically seen in most assessments unless extended performance tasks are included.

Table IV-20 shows the percentage of operational items by DOK level, subject, and grade. This information also reveals the proportions of DOK requirements according to content standards. Most ELA items are at Level 1 and Level 2; fewer items are at Level 3. In mathematics, most

items are at Level 1 and Level 2 as well, with relatively fewer Level 3 items. For science, most items are at Levels 2 and 3, with a few items at Level 1.

Table IV-20. Number of Items by DOK Level, Subject, and Grade

	ELA				Math				Science			
	DOK level, %				DOK level, %				DOK level, %			
Grade	Total items	1	2	3	Total items	1	2	3	Total items	1	2	3
3	86	25	54	7	88	34	52	2				
4	82	20	52	10	80	25	50	4				
5	81	23	46	12	87	34	53	0	42	6	20	16
6	85	32	40	13	85	34	50	1				
7	74	9	52	13	80	35	43	2				
8	80	20	51	9	83	24	54	5	48	4	23	21
10	68	16	48	4	83	28	49	6				
11									47	1	24	20

IV.4. Scoring and Scaling

This section discusses the procedures of scoring individual items, scoring the test as a whole, and scaling.

IV.4.1. Item scoring. KAP assessment items administered in 2017 are all machine scored. The online test delivery platform compares student responses to the correct keys stored with the items and assigns the predetermined scores accordingly.

IV.4.2. Test scoring. Test scoring uses a psychometric model to derive item scores on the test to produce a single score indicating a student's proficiency level. The IRT ability estimates (thetas) are computed using the 2PL model and GRM. Because the total score is derived using the number-correct method—in which scores for each item are added together to derive the raw score—thetas have one-to-one correspondence with raw scores (i.e., each raw score has only one matching theta). Using the test characteristic curve function of the IRT models, the theta for each raw-score point is obtained for a test form (Press, Flannery, Teukolsky, & Vetterling, 1989).

IV.4.3. Scaling. Scaling is the procedure of transforming thetas or raw scores to a scale. The purpose is to facilitate the use and interpretation of test scores. The scale is also the basis for setting performance levels. The theoretical values of theta range from negative infinity to positive infinity. In other words, thetas can be negative values and have decimal points. One can imagine the difficulty of using and interpreting negative test scores with multiple decimal points. To ease score interpretation, it is crucial to transform thetas to a scale composed of positive integers.

The section below addresses the procedures for constructing scaled scores. Procedures used to establish ELA and mathematics performance-level cut scores can be found in the 2015 Technical

Manual. Procedures used to establish science cut scores are described in the current manual in VI: Academic Achievement Standards and Reporting.

IV.4.3.1. Scale transformation and cut scores. Kolen and Brennan (2004) used the following formula to derive scaling constants:

$$SS(y) = \frac{\sigma(SS)}{\sigma(Y)} y + [SS(y_1) - \frac{\sigma(SS)}{\sigma(Y)} y_1], \qquad \text{(IV-3)}$$

where $SS(y)$ is the scaled score, $\sigma(SS)$ is its *SD*, $\sigma(Y)$ is the *SD* of the original scores, y_1 is an original score, and $SS(y_1)$ is the scaled score equivalent to the original score, y_1. This equation can be structured to

$$SS = A \times y + C, \text{ where} \qquad \text{(IV-4)}$$

$$A = \frac{\sigma(SS)}{\sigma(Y)} \text{ and} \qquad \text{(IV-5)}$$

$$C = SS(y_1) - \frac{\sigma(SS)}{\sigma(Y)} y_1. \qquad \text{(IV-6)}$$

A and *C* are the slope and intercept of the scaling constants, respectively. KSDE has predetermined the scaled score to have a slope, *A*, of 25 for all subjects and grades.

The KAP assessment has four performance levels, Level 1 through Level 4; achieving higher levels indicate higher performance. Students in Levels 3 or 4 are considered proficient. A scaled score of 300 is determined by KSDE as the cut that separates Levels 2 and 3 (Level 2/3). In other words, a scaled score of 300 separates students into proficient and nonproficient groups. The original theta values of Level 2/3 cuts of each subject and grade were set by standard-setting panels. With the original cut score (y_1), equivalent scaled score (i.e., $SS(y_1) = 300$), and a scaled-score *SD* of 25 (i.e., $\sigma(ss) = 25$) identified, the intercept, *C*, can be derived using Equation IV-6 after the *SD*, $\sigma(Y)$, is computed.

IV.4.3.2. ELA and mathematics scale transformation. Equating of ELA and mathematics is conducted with IRT models; thus, their initial ability estimates are the IRT thetas. Because thetas are used for ELA and mathematics, the y_1 in Equation IV-6 is the theta associated with a scaled score of 300. The grade-level theta cuts for ELA and mathematics were set by standard-setting panels in 2015 (see theta cuts in Tables IV-21 through IV-23). Using Equation IV-6, the *C* for each grade is found (see Table IV-24). Because *A* and *C* are known, the other two scaled-score cuts can be derived using Equation IV-4. Note that the scaled-score cuts are rounded up rather than to the nearest integer. The rationale for rounding up is that students need to have scores equal to or higher than the cut score to pass a given level.

Table IV-21. ELA Cut Scores

Grade	Theta cuts			Scaled-score cuts		
	Level 1/2	Level 2/3	Level 3/4	Level 1/2	Level 2/3	Level 3/4
3	−1.015	−0.050	1.020	276	300	327
4	−1.457	−0.275	1.107	271	300	335
5	−1.085	−0.064	0.952	275	300	326
6	−0.756	0.181	1.594	277	300	336
7	−0.800	0.219	1.610	275	300	335
8	−0.940	0.495	1.850	265	300	334
10	−0.785	0.465	1.800	269	300	334

Table IV-22. Mathematics Cut Scores

Grade	Theta cuts			Scaled-score cuts		
	Level 1/2	Level 2/3	Level 3/4	Level 1/2	Level 2/3	Level 3/4
3	−1.225	−0.230	0.906	276	300	329
4	−1.215	0.160	1.375	266	300	331
5	−0.885	0.219	1.245	273	300	326
6	−0.882	0.215	1.340	273	300	329
7	−1.055	0.321	1.980	266	300	342
8	−0.527	0.530	1.968	274	300	336
10	−0.497	0.530	1.830	275	300	333

Table IV-23. Science Cut Scores

Grade	Theta cuts			Scaled-score cuts		
	Level 1/2	Level 2/3	Level 3/4	Level 1/2	Level 2/3	Level 3/4
5	−0.940	−0.030	1.160	277	300	330
8	−0.600	0.400	1.505	275	300	328
11	−0.550	0.315	1.450	278	300	328

Table IV-24. ELA, Mathematics, and Science Scaling Constants

	ELA		Mathematics		Science	
Grade	*A*	*C*	*A*	*C*	*A*	*C*
3	25	301.25	25	305.75		
4	25	306.87	25	296.00		
5	25	301.59	25	294.53	25	300.75
6	25	295.48	25	294.63		
7	25	294.53	25	291.98		
8	25	287.63	25	286.75	25	290.00
10	25	288.38	25	286.75		
11					25	292.13

Note. *A* = slope; *C* = intercept.

IV.4.3.3. Properties of scaled scores. The derived scaled scores are decimal numbers and must be rounded up to the nearest integers. The IRT model cannot estimate the thetas of extreme scores (e.g., 0 and perfect raw scores) because responses to all items are identical. A theta of −99 or 99 is typically assigned to those raw-score points by software. To keep the scaled score meaningful, the lowest obtainable scaled score (LOSS) and the highest obtainable scaled score (HOSS) are set to cap scaled scores within a reasonable range. KAP's LOSS and HOSS are set as 220 and 380, respectively.

IV.4.4. Operational test results. Summaries of scaled scores by subject and grade are presented in Tables IV-25 through IV-27; summaries by demographic subgroups are presented in Appendix D. Graphs of scale-score distribution by subject by grade can be found in Appendix F. Tables IV-25 through IV-27 indicate that the minimum and maximum values are within the LOSS and HOSS values of 220 and 380, respectively. The differences between (a) P_{50} and P_{25} and (b) P_{75} and P_{50} are indicators of the shapes of score distributions: The larger of the two differences indicates the direction of any skewness 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 match, the distribution is symmetric. In ELA, the distributions for grades 3 and 4 are symmetric in shape; the distributions for grades 6 and 10 are negatively skewed; and the distribution for grades 5, 7, and 8 are positively skewed. In mathematics and science, all distributions are positively skewed.

Table IV-25. Scaled-Score Descriptive Statistics by Grade for ELA

Grade	*M*	*SD*	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
3	295.4	28.8	220	259	274	295	316	335	380
4	300.5	28.0	220	265	279	300	321	338	380
5	297.0	29.9	220	257	275	295	317	335	380
6	291.1	28.9	220	252	270	292	312	327	380
7	289.7	30.7	220	252	268	287	310	330	380
8	284.0	28.4	220	247	263	282	304	321	380
10	284.8	29.8	220	247	263	284	304	325	380

Note. P_{10}, P_{25}, P_{50}, P_{75} and P_{90} are 10th, 25th, 50th, 75th, and 90th percentiles, respectively.

Table IV-26. Scaled-Score Descriptive Statistics by Grade for Mathematics

Grade	*M*	*SD*	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
3	303.2	27.5	220	268	284	302	321	339	380
4	293.9	28.2	220	260	273	291	312	332	380
5	291.1	27.6	220	258	271	288	308	329	380
6	291.3	26.8	220	261	271	287	306	327	380
7	288.4	27.7	220	255	269	284	305	327	380
8	284.7	29.1	220	253	265	280	300	325	380
10	285.6	28.5	220	256	265	279	299	325	380

Note. P_{10}, P_{25}, P_{50}, P_{75} and P_{90} are 10th, 25th, 50th, 75th, and 90th percentiles, respectively.

Table IV-27. Scaled-Score Descriptive Statistics by Grade for Science

Grade	*M*	*SD*	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
5	298.5	30.2	220	260	277	297	318	338	380
8	288.4	29.6	220	252	268	286	306	327	380
11	291.7	29.1	220	257	270	289	309	330	380

Note. P_{10}, P_{25}, P_{50}, P_{75} and P_{90} are 10th, 25th, 50th, 75th, and 90th percentiles, respectively.

The scaled-score means presented in Appendix D show that across all subjects and grades, Asian students have the highest mean scores, followed by White students. African American students do not perform as well as other groups do. The gaps between the highest and smallest subgroup mean scores range between 25 and 30 scaled-score points for ELA, 29 and 39 for mathematics, and 22 and 27 for science. In terms of *SD*s of scaled scores, most mean score differences between groups fall within the range of 1 *SD*. Some exceptions are as follows:

- ELA: mean score difference between students with and without disabilities is slightly greater than 1 *SD* in grades 6, 7, 8, and 10;
- Mathematics: mean score difference between students with and without disabilities is slightly greater than 1 *SD* in grade 7;
- Science: mean score difference between English and non-EL is slightly greater than 1 *SD* in grade 8.

The proportion of students in each performance level (Levels 1 through 4) and college- and career-ready rate (combined Levels 3 and 4) are provided by subject and grade in Table IV-28 and Figures IV-1 through IV-3. The readiness rates ranged from 25% to 54% across subjects and grades. All three subjects tended to have lower readiness rates in higher grade levels.

Table IV-28. Percentage of Students in Each Performance Level by Subject and Grade

	ELA (%)					Mathematics (%)					Science (%)				
Grade	1	2	3	4	CCR	1	2	3	4	CCR	1	2	3	4	CCR
3	27	31	28	14	**42**	16	29	37	17	**54**					
4	15	35	39	11	**50**	16	44	30	11	**40**					
5	24	30	29	17	**46**	28	38	22	12	**34**	24	30	31	15	**46**
6	32	27	35	5	**40**	28	39	24	10	**33**					
7	33	32	26	9	**35**	21	49	26	4	**30**					
8	27	45	23	5	**28**	40	34	20	6	**26**	34	32	24	10	**34**
10	31	38	25	6	**30**	43	33	17	8	**25**					
11											36	27	24	13	**37**

Note. College- and career-ready (CCR) rates are in boldface.

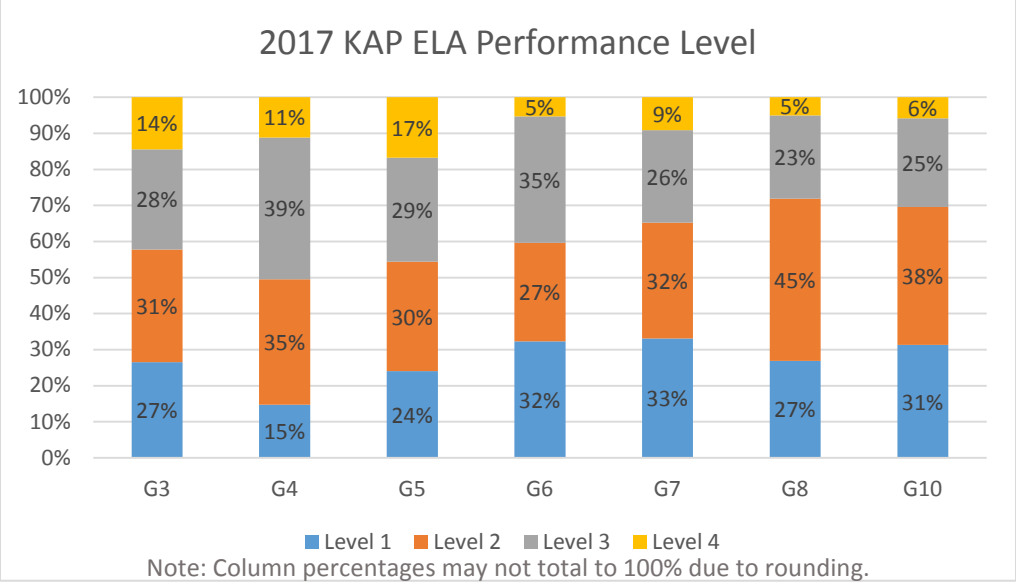


Figure IV-1. Performance-level results for ELA.

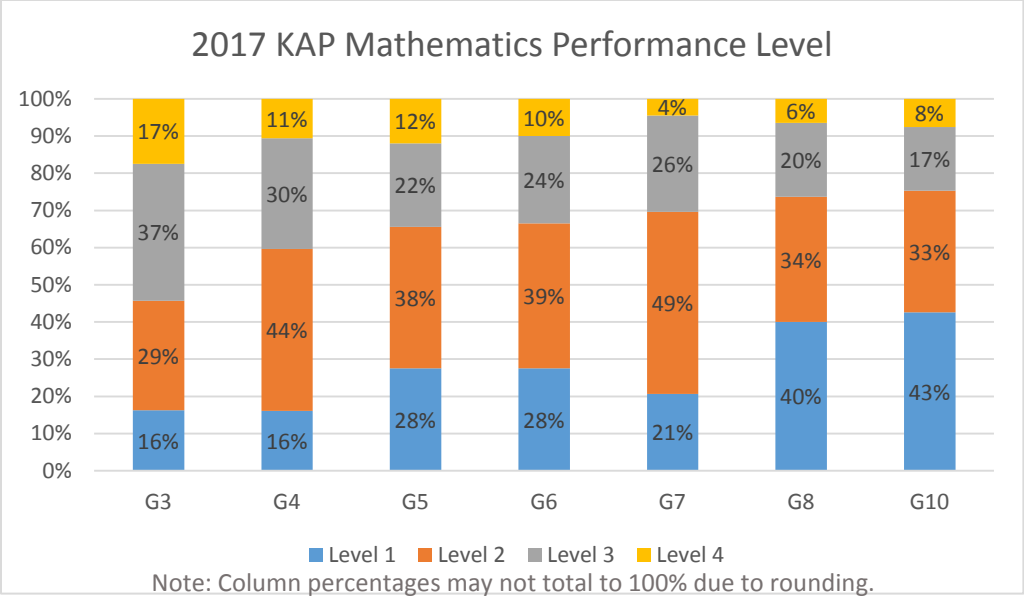


Figure IV-2. Performance-level results for mathematics.

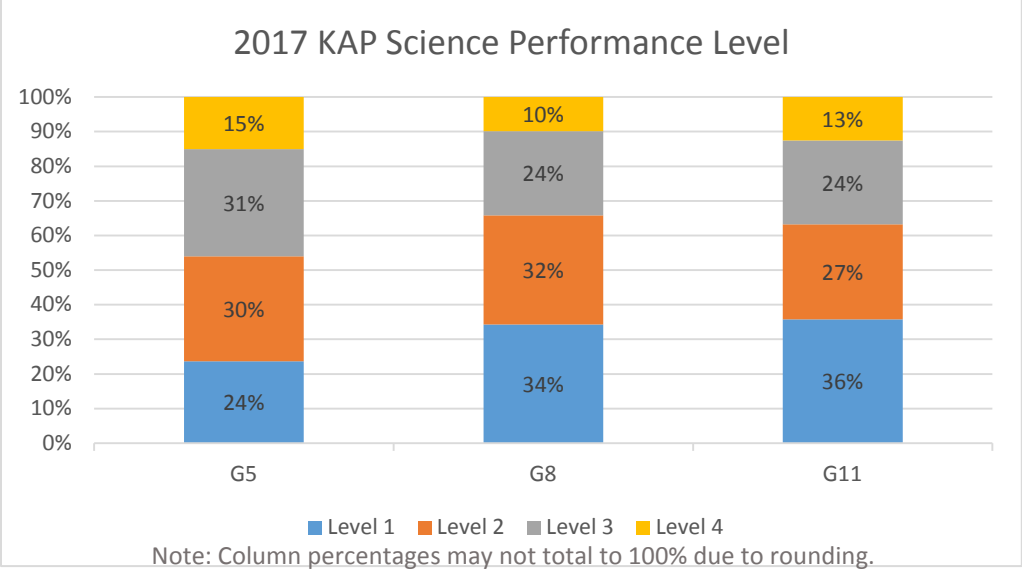


Figure IV-3. Performance-level results for science.

ELA and mathematics scaled-score and performance-level trends across years are presented in Tables IV-29 and IV-30 and in Figures IV-4 through IV-7. The tables present the scaled-score mean, *SD*, and *N* count across administration years by grade. Figures IV-5 and IV-6 present the trends by performance level, and Figures IV-6 and IV-7 present trends in the percentage of student scores in Levels 3 and 4. The longitudinal trend cannot be computed for science because 2017 is the first year of its administration. Figures IV-6 and IV-7 show that ELA Level 3/4 percentages declined in all grades; mathematics Level 3/4 percentages increased in grades 4 and 10 but decreased in grades 3, 5, 6, and 8.

Table IV-29. Longitudinal Scaled-Score Trend for ELA

	2015			2016			2017		
Grade	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*
3	298.3	24.7	37,723	298.7	28.0	38,370	295.4	28.8	38,340
4	303.6	24.9	37,200	303.0	29.3	37,366	300.5	28.0	38,424
5	298.6	25.0	36,965	298.2	29.3	36,803	297.0	29.9	37,526
6	292.9	24.7	37,270	293.1	28.4	36,732	291.1	28.9	36,858
7	291.3	25.1	36,875	292.7	27.9	36,589	289.7	30.7	36,863
8	285.9	24.7	36,784	286.7	28.5	36,193	284.0	28.4	36,695
10	286.7	24.7	35,593	285.3	29.6	35,653	284.8	29.8	35,673

Table IV-30. Longitudinal Scaled-Score Trend for Mathematics

	2015			2016			2017		
Grade	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*
3	303.2	24.4	37,740	304.3	27.8	38,343	303.2	27.5	38,438
4	293.0	24.7	37,261	293.3	28.2	37,448	293.9	28.2	38,514
5	292.2	24.5	36,986	292.2	27.4	36,806	291.1	27.6	37,608
6	292.6	23.9	37,268	292.2	27.2	36,657	291.3	26.8	36,923
7	289.6	24.0	36,878	289.4	28.4	36,583	288.4	27.7	36,910
8	285.7	23.9	36,821	285.5	28.4	36,169	284.7	29.1	36,758
10	285.7	23.7	35,603	285.0	28.2	36,831	285.6	28.5	35,653

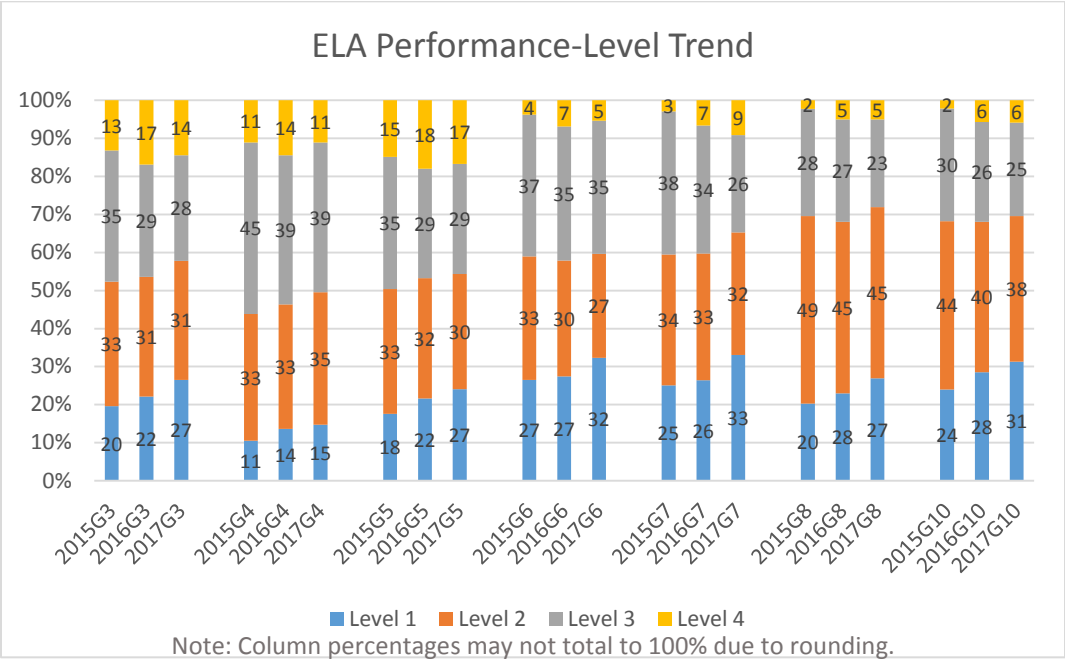


Figure IV-4. Performance-level trend for ELA. G = grade.

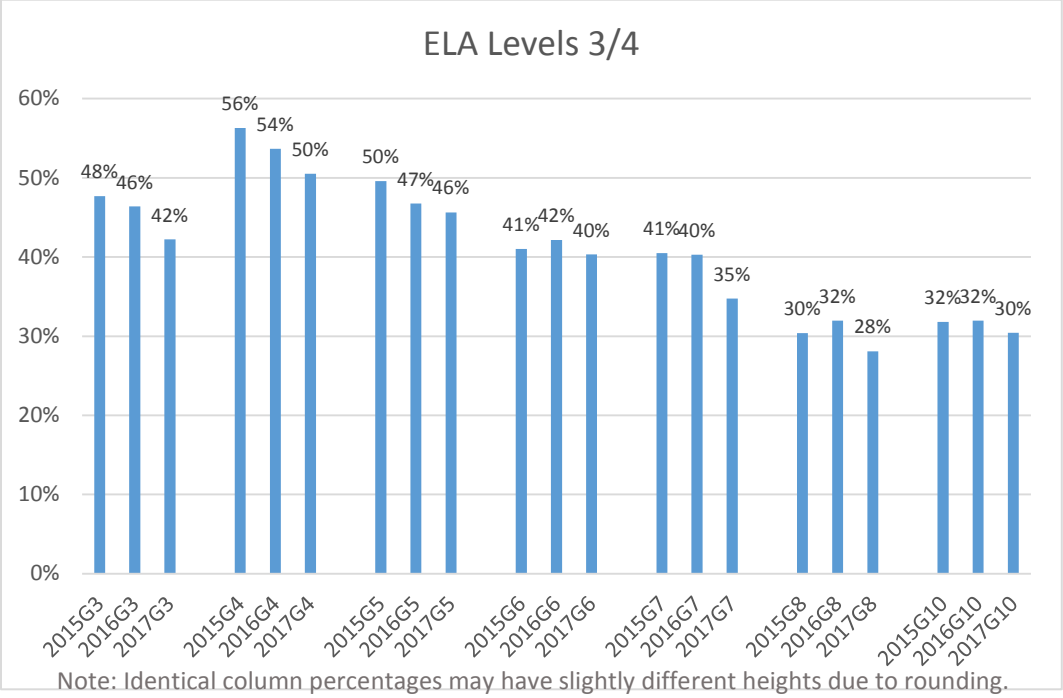


Figure IV-5. Career-readiness trend for ELA. G = grade.

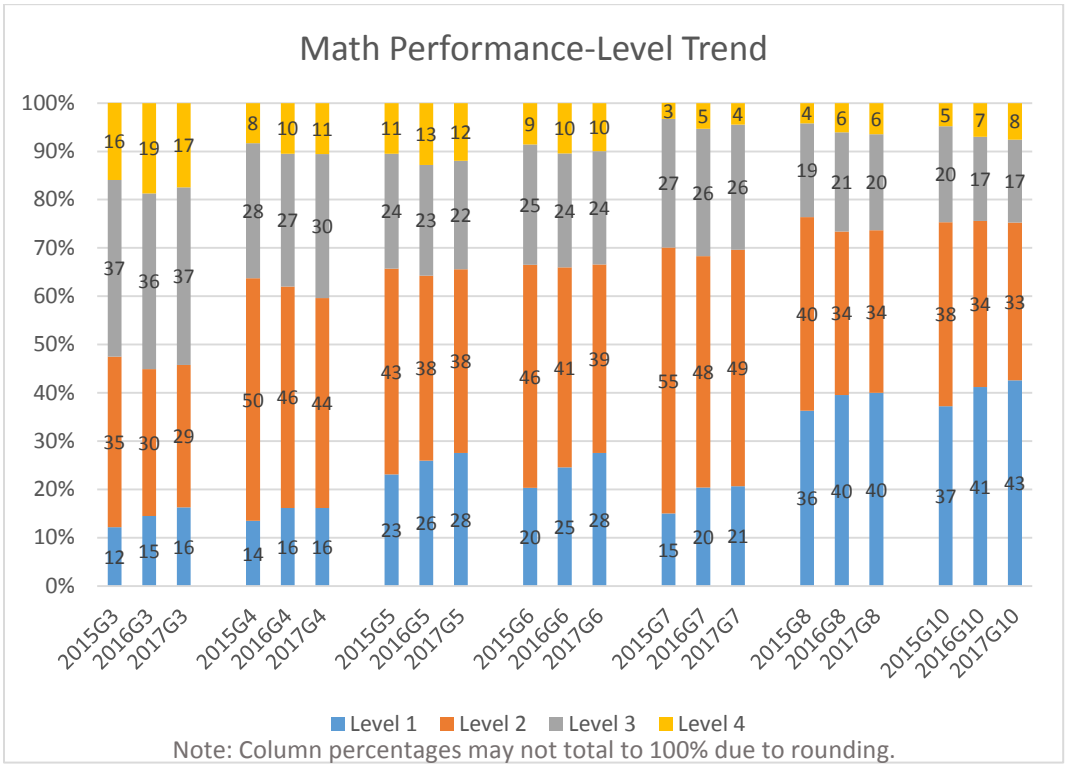


Figure IV-6. Performance-level trend for mathematics. G = grade.

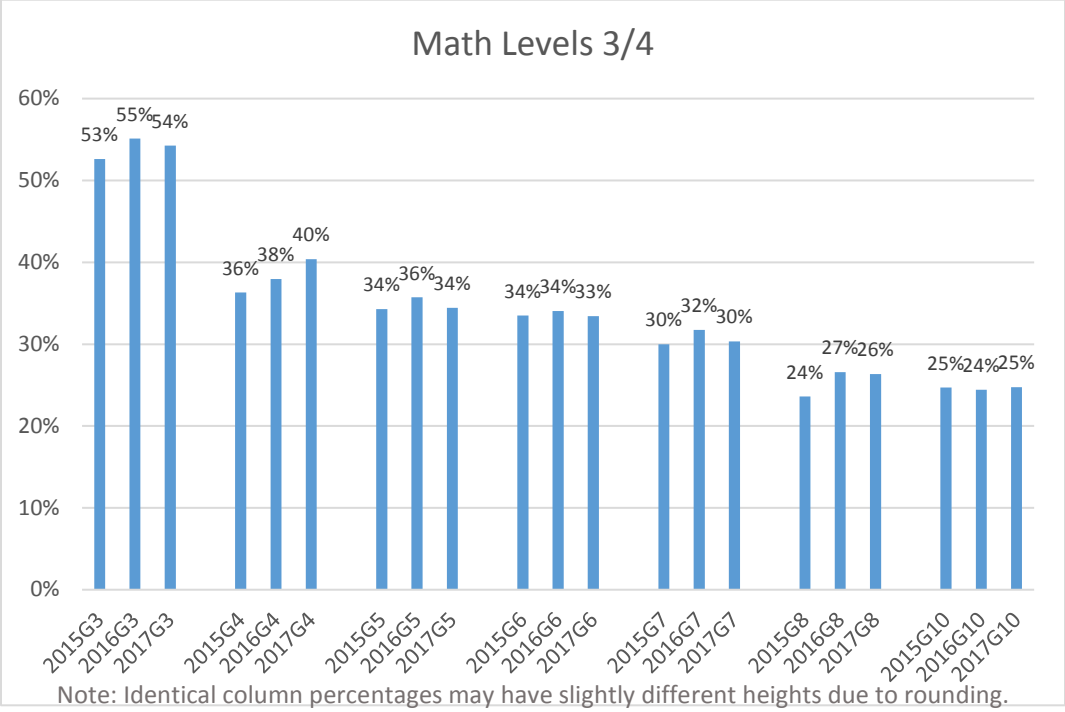


Figure IV-7. Career-readiness trend for mathematics. G = grade.

IV.5. Multiple Assessment Forms

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

The 2017 KAP administration is in its third year of KCCRS administration for ELA and mathematics. Both subjects have multiple operational forms, so their linking involves both within- and cross-year equating procedures. The first year, 2016, of KCCRS administration for science and each grade has multiple operational forms. Thus, only the within-year linking was conducted.

IV.5.1. Within-year linking design. The within-year linking for ELA and mathematics tests is achieved through pre-equating. All items on the ELA and mathematics tests have been placed on the same IRT scale using 2015 and 2016 calibrations. All items on the science tests were calibrated together using concurrent item calibration and are placed on the same IRT scale. When the items from different test forms are on the same IRT scale, the student scale scores calculated through these IRT item parameters are equated.

IV.5.2. Cross-year linking design. To increase the number of linking items and maximize

linking stability, the cross-year linking uses the pre-equating method. All items on the 2017 ELA and mathematics tests have IRT parameters calibrated in previous years, and they are on the same IRT scale as items in 2015 and 2016 tests. When the items from different years are on the same IRT scale, the student scale scores calculated from these IRT item parameters are equated and placed onto the base scale (i.e., the 2015 scale).

IV.5.3. Linking procedure. The concurrent item calibration for science was conducted using flexMIRT (Cai, 2013; Houts & Cai, 2013). Refer to section III.3.2. for details.

IV.6. Technical Analysis and Ongoing Maintenance

For the 2017 assessment, KSDE requested a more compact test, reducing the amount of instructional time devoted to testing. Changes to the test included a change in the adaptive model, elimination of performance tasks in math and on-demand writing in ELA, and removing items from the ELA test that measured listening. KSDE received a waiver from the US Department of Education to not require the assessment of Speaking and Listening from the ELA standards.

IV.6.1. Model change. The prior stage-adaptive model was a 1–3–4 model with a separate, non-adapting block for field test items. To shorten the tests, the model was changed to a 1–2 model, with field test items embedded in the first (non-adapting) stage.

Multiple Stage 1 blocks were created, which varied by the content of the field test items. Additionally, a parallel and equivalent Stage 1 block and a set of field test items were created as an accessible Stage 1. Students requiring accommodations, such as text-to-speech, were administered the accessible Stage 1 block. All other students were randomly assigned to either the accessible Stage 1 block or one of the other Stage 1 blocks that did not have additional accessibility features. The Stage 1 block was developed to have a moderate theta to cover a wide range of item difficulty to serve its purpose as a routing test,and to cover the entirety of the content standards so that the routing decision would be based on a sample of content from the entire domain of ELA or mathematics.

Stage 2 routed into harder (higher theta) or easier (lower theta) blocks. Parallel and equivalent accessible Stage 2 blocks were created for both the harder and easier block. Test takers were routed into a Stage 2 block based on the person-estimate coming out of the operational items in the Stage 1 block. Field test items did not count for or against the student and did not calculate in either the routing person-estimate or the final proficiency estimate from the entire test.

Shortening the test and reducing the number of paths had a positive result of allowing test developers to select stronger items for the test. The original 1–3–4 model required a minimum of 135 items to complete an adaptive panel (not taking into account linking items or accessibility considerations). The more parsimonious 1–2 model required 85 items for mathematics and 72 items for ELA. Passage needs for ELA were also reduced from nine passages in the fuller model to six passages in the more parsimonious model.

IV.6.2. Elimination of hand-scored tasks. The 2016 mathematics assessment included a performance task that was an extended problem set couched in a real-world scenario. The mathematics performance task was generally given on a separate day from the machine-scorable test. The 2016 ELA test included an on-demand writing task that involved a testing period of reading and note-taking followed by a testing period of writing on the same day or the following day. The time and expense involved in administering an additional 1- or 2-day test that required human scoring led KSDE to decide to abandon the performance component of both the ELA and mathematics assessment.

Because the score of the ELA on-demand writing task was not counted toward the scale score in the 2016 test but was combined with the scale-score performance level to generate a combined performance level, removing the ELA on-demand writing task in 2017 will not have any effect on the scale score. The 2016 mathematics performance tasks, however, were included in calculating scale scores. The effect of removing those tasks on scale-score and performance-level classification were studied using the 2016 data set. The comparison of performance-level classification, scale-score statistics, and reliability between two sets of scale scores and impact data derived from the 2016 data indicated there almost no differences occur between the test scores with performance tasks included and without them included. Table IV-31 presents mathematics scale scores from the 2016 test in which performance tasks were included and from the 2017 test in which performance tasks were excluded. Nearly no differences occur between the two sets of the scores. Thus, elimination of the performance component did not affect the classification consistency or test scores. Table IV-32 shows reliability differences, if any, between mathematics tests with and without performance tasks.

Table IV-31. Mathematics Scale Scores by Grade: A Comparison Between Tests with and Without Performance Tasks

Grade	With PT *M*	Without PT *M*	With PT *SD*	Without PT *SD*
3	304.4	304.6	27.8	27.7
4	293.3	293.4	28.2	28.4
5	292.3	292.4	27.4	27.5
6	292.2	292.4	27.2	27.5
7	289.4	289.6	28.4	28.7
8	285.6	285.8	28.4	28.8

Note. PT = performance task.

Table IV-32. Reliability Difference Between Mathematics Tests with and Without Performance Tasks

Grade	With PT reliability	Without PT reliability
3	0.94	0.93
4	0.95	0.95
5	0.95	0.94
6	0.94	0.93
7	0.93	0.92
8	0.94	0.93

Note. PT = performance task.

IV.6.3. Elimination of Listening. With the receipt of a waiver from the US Department of Education relaxing the requirement that states assess Speaking and Listening components of ELA, KSDE requested that the 2017 assessment not include the six planned items that would measure listening. A study was conducted to compare the performance-level classification, scale-score statistics, and reliability between two sets of scale scores and impact data derived from the 2016 test data: a dataset including listening items and the other without listening items. Results of this comparison indicated that differences on these statistics are very small. Table IV-33 presents ELA scale scores from the 2016 test in which listening items were included and from the 2017 test in which listening items were excluded. Nearly no differences occur between the two sets of the scores. Thus, elimination of listening items did not affect the classification consistency and test scores. Table IV-34 shows reliability differences, if any, between ELA tests with and without performance tasks.

Table IV-33. ELA Scale Scores by Grade: A Comparison Between Tests with and Without Listening Items

Grade	With listening *M*	Without listening *M*	With listening *SD*	Without listening *SD*
3	298.7	298.8	28.0	28.0
4	303.0	302.6	29.3	29.2
5	298.2	298.3	29.3	29.3
6	293.1	293.2	28.4	28.7
7	292.8	292.9	27.8	28.2
8	286.8	286.8	28.5	28.7
10	285.3	285.3	29.6	29.6

Table IV-34. Reliability Difference Between ELA Tests with and Without Performance Tasks

Grade	With listening items	Without listening items
3	0.92	0.92
4	0.91	0.90
5	0.91	0.91
6	0.91	0.91
7	0.90	0.89
8	0.91	0.91
10	0.92	0.92

While a shorter test permitted the careful selection of the strongest items, ongoing maintenance and refreshing of item pools is a vital part of maintaining the validity and credibility of an assessment program.

V. Inclusion of All Students

KSDE complies with the Individuals with Disabilities Education Improvement Act (IDEA) and the Elementary and Secondary Education Act (ESEA), both of which require that all students, including students with disabilities, participate in assessments used for accountability purposes. One of the basic reform principles of ESEA is stronger accountability for educational achievement results for all students. Through this federal legislation, assessments that aim to increase accountability provide important information regarding (a) schools' success in including all students in standards-based education, (b) students' achievement of standards, and (c) improvements needed for specific groups of students. IDEA explicitly governs services provided to students with disabilities. Accountability at the individual level is provided through the Individualized Education Program (IEP) developed to address each student's unique needs.

Assessment accommodations are practices and procedures that provide equitable access during instruction and assessments for students with special needs. These accommodations may not alter the assessment's validity, score interpretation, reliability, or security. They are intended to reduce or eliminate the effects of a student's disability; however, they do not alter learning expectations. The accommodations provided to a student should be the same for classroom instruction, classroom assessments, and local educational agency and state assessments. It is critical to note that some accommodations that are appropriate for instructional uses may not be appropriate for use on standardized assessments. For example, a student with low vision will need accommodations to make a test accessible. However, in an ELA assessment, reading passages aloud to a student would change what is being measured and therefore is not a valid accommodation. Use of a magnifying tool or a large-print version of a test is an acceptable accommodation, though. It is important for educators to become familiar with state policies regarding accommodations during assessments.

This chapter presents information about KAP's inclusion of all students and accommodation usage. Much of this information is also available in other KSDE documents (e.g., *Tools and Accommodations for the Kansas Assessment Program (KAP) 2017* and the *Kansas Assessment Examiner's Manual 2016–2017*). This chapter closes with a report of the frequency of use of specific accommodations.

V.1. Procedures for Including Students with Disabilities

KSDE is committed to including all students in the KAP assessments. The inclusion of students with disabilities is achieved by providing clear guidelines for educators to register their students with different needs. The *Examiner's Manual* describes step-by-step registration procedures for students who need accommodations. Additionally, educators are instructed to report students who are not assessed. Some notable exceptions that occur in Kansas include the following:

- students serving a long-term suspension,
- students who were truant more than two consecutive weeks at time of testing,
- students who had catastrophic illness or accidents,
- students who moved during testing, or
- students who were incarcerated.

V.2. Accommodations

A few basic rules apply to every available accommodation on the KAP assessment. First and foremost, only accommodations that have been used regularly in instruction may be used on state assessments. Second, students with IEPs, 504 plans, or EL plans, as well as students with Student Improvement Team plans, may use only the accommodations documented in their plans. Finally, for accommodations to be available during the KAP assessment, teachers must submit accommodation requests through the student's PNP in Educator Portal.

Test Administrators handle some accommodations that are allowed for the KAP assessment, but some accommodations are built-in features in the KITE system. Because features in the KITE system are activated according to students' needs, teachers are required to mark those needs in the PNP. Additionally, teachers need to report in advance if braille is needed. Table V-1 shows available accommodations according to reporting requirement.

Table V-1. Available Nonreported and Reported KAP Accommodations

Nonreported	Reported
Allowable practice	Auditory background
Delivery of directions to student in ASL	Background color
Frequent breaks	Braille
Separate, quiet, or individual setting	Color overlay
Spanish translation	Foreground color
Student dictation of answers to scribe	Magnification
Student reading assessment aloud to self (via headset)	Invert color choice
Student response in American Sign Language (ASL)	Item-translation display
Student use of braille writer or slate and stylus	Keyword-translation display
Student use of communication device	Large-print booklet
Student use of translation dictionary	Masking
Text-to-speech	Onscreen keyboard
Use of some other accommodation	Signing
	Speech (read aloud)
	Touch

V.3. Frequency of Accommodation Use

The PNPs submitted by teachers determine the availability of online test accommodations for individual students. Thus, the summary of PNP accommodation requests shown below also indicate the number of students for whom each accommodation is requested. Table V-2 summarizes the PNPs by subject and grade; note that some students may receive multiple accommodations. The table shows that "Spoken (read aloud)" is the most commonly used accommodation option.

Table V-2. Frequency of Accommodation Requests by Grade

	Grade							
Accommodation	3	4	5	6	7	8	10	11
Auditory background	40	51	77	82	32	36	26	32
Background color	54	76	55	71	61	66	61	47
Braille	4	7	4	7	5	9	4	5
Color overlay	35	71	65	89	56	49	45	35
Foreground color	54	76	55	71	61	66	61	47
Invert color choice	21	34	21	32	24	30	22	17
Item-translation display	0	12	24	29	40	39	27	27
Keyword-translation display	97	139	202	184	235	264	91	76
Large-print booklet	0	0	11	15	9	8	10	2
Magnification	114	133	113	115	100	120	97	85
Masking	6	8	2	17	1	1	1	1
Onscreen keyboard	23	24	36	30	40	29	50	33
American Sign Language	13	11	13	15	7	11	14	11
Spoken (read aloud)	4,219	4,590	4,531	4,137	3,887	3,402	2,467	1,991
Tactile	0	0	1	3	0	0	2	3
Touch	9	7	16	5	8	4	9	5
Total	4,689	5,239	5,226	4,902	4,566	4,134	2,987	2,417

VI. Academic Achievement Standards and Reporting

VI.1. State Adoption of Academic Achievement Standards for All Students

PLDs define the KAP academic achievement standards. While a test is developed according to content standards, students' performances are evaluated using the academic achievement standards. PLDs describe the expected academic performances at each performance level. When a performance level is assigned to a student, the student meets the minimum expected knowledge and skills of that performance level. This score interpretation applies to all students who participated in the KAP assessment.

VI.2. Achievement Standard Setting

ELA and mathematics standard setting occurred in 2015. The procedures and outcomes can be found in the 2015 technical manual. The HGSS standard setting occurred in 2016, and the procedures and outcomes can be found in the 2016 technical manual. This section focuses on science standard setting. CETE conducted the standard setting for science using the Bookmark method during a workshop held at a school in Topeka, Kansas, on June 20–21, 2017. The standard-setting event included a training session and three rounds of the modified Bookmark procedure for each grade/subject area test. The main goal of the science standard setting was to establish three cut scores that differentiate four performance levels for the assessment. The panelists' recommended cut scores were presented to the State Board for approval.

VI.2.1. Overview of the Bookmark method. The standard Bookmark procedure (Mitzel, Lewis, Patz and Green, 2001) is a complete set of activities designed to generate cut scores based on panelists' reviews of collections of test items (Cizek & Bunch, 2007). In this method, items are presented in an ordered item booklet (OIB) from easiest to hardest based on empirical item data (i.e., IRT item-parameter estimates). Panelists are asked to review these items and to express their judgments by placing a bookmark at the page in the OIB where they believe the just-barely examinee would not have a specific probability (i.e., 67%) of answering the item correctly.

The Bookmark method capitalizes on the fact that IRT scaling places both items and students on the same scale. Given that the assumptions of the IRT model hold, a student's test score can provide a theoretically known probability for the student answering a given multiple-choice item correctly, or in the case of polytomously scored items (e.g., constructed-response (CR) items), obtaining a given score point.

According to Cizek & Bunch (2007), the Bookmark procedure has become quite popular for several reasons. First, from a practical perspective, the method can be used for complex, mixed-format assessments, and panelists using the method consider selected-response (SR) and CR items together. Second, from the perspective of those who will be asked to make judgments, it presents a relatively simple task to participants. Third, in addition to being easy for participants, the Bookmark method is also comparatively easy for those who must implement the procedure.

Finally, from a psychometric perspective, the method has certain advantages because of its basis in IRT analysis and because of the fidelity of the method to the test construction techniques that were used in the development the assessment. Given that the KAP assessment had a suitably large number of items in its pool that adequately covered the achievement range of all performance levels and that the tests were taken by a reasonably large number of students, KSDE and CETE believed the item data would be appropriate for the application of the Bookmark method.

VI.2.2. The ordered item booklet (OIB). The OIB can contain both dichotomously scored items (e.g., multiple choice) or polytomously scored items (e.g., items with partial credit scoring) intermingled in the same booklet. A dichotomously scored item appears in the OIB once, in a location determined by its difficulty (usually its IRT *b* value). A polytomously scored item appears several times in the booklet, once for each of its score points. Each dichotomous item will have one associated difficulty index, and each polytomous item will have as many step (difficulty) functions as it has score points (excluding zero).

The OIB can comprise any collection of items spanning the range of content, item types, and difficulty represented in a typical test and need not consist only of items that have appeared in an intact test. Therefore, the OIB can have more or fewer items than an operational test form. By permitting items beyond those included in an operational test form, the gaps in item difficulty or content coverage can be filled with items from a bank. For the science OIBs, items were selected so that no noteworthy gaps in item difficulty or content overage existed.

VI.2.3. Panelist recruiting process. The selection and training of the standard-setting panelists are crucial to the success of a standard-setting event. Considering several aspects of panel diversity (e.g., ethnicity, gender, geographic area, teaching experience, and role), KSDE took several steps to recruit panelists that represent the variety of the Kansas educator population for the standard-setting workshop. To obtain a large and diverse pool of applicants, KSDE began recruitment efforts early in the year. Invitations were sent to all teachers and administrators in the current educator database, and the invitation was extended to those educators' colleagues in case some educators were not in the database. Additional recruitment efforts were also made through relationships with school district and individual educators. When selecting panelists from the applicant pool, KSDE reviewed all applications and placed emphasis on ethnic, gender, and geographic diversity.

KSDE also gave first preference to teachers who did not participate in item reviews or the PLD committee. Other factors considered in panelist selection included current licensure type, content endorsements, and EL or special education endorsements. Namely, the selected panelists should represent the following:

- All 10 State Board districts,
- Priority/focus schools,
- A cross-section of state large/small districts, rural/urban districts, and socioeconomic composition of districts, and

- A range of length of teaching experience (i.e., new/veteran teacher).

VI.2.4. Performance level descriptors (PLDs). As mentioned previously, PLDs describe the expected academic performance standards at each performance level. Thus, PLDs are the guiding performance standards when setting cut scores. The creation of science PLDs began with KSDE and CETE content staff, who developed descriptors for the content that all students should know and be able to achieve at each performance level. These descriptors adhered to the cognitive alignment of the content standards, such as DOK, cognitive complexity, scope of skills, inquiry vs. process, etc. (see Appendix H). KSDE staff and Kansas educators reviewed and approved the grade-specific PLDs for all four levels prior to the standard-setting workshop.

VI.2.5. Standard-setting procedure. The standard-setting activities described in this section follow the event in chronological order, as laid out by the meeting agenda (see Appendix I). Each grade had three panels, and each panel had three to five panelists. For each grade level, the standard-setting procedures were steered by a lead facilitator and three table leads recruited by CETE. Both KSDE assessment personnel and CETE content team members were available at the workshop to address policy- or content-related questions. A description of the workshop structure follows.

June 20

- Completed the training session
- Competed the science exam and reviewing items
- Completed the "just-barely" student activity
- Practiced bookmarking
- Wrote item knowledge and skills for test items on OIB

June 21

- Completed the readiness form
- Completed three rounds of bookmarking
- Completed evaluation form
- Completed training on articulation
- Reviewed and discussed Round 3 results
- Articulated cut scores across grades as a group
- Completed articulation evaluation form

VI.2.5.1. Training session. At the start of the meeting, panelists completed a participant survey form (see Appendix J) and signed a confidentiality form (see Appendix K). The survey collected panelist biographical data to contribute to the documentation of the procedural validity of the standard-setting process (Hambleton, Pitoniak, & Copella, 2012; Pitoniak & Morgan, 2012; Rosseel, 2012). Then, CETE staff conducted a large group training to address general topics that included an overview of the science assessment and an introduction to the concept of cut scores. CETE staff also introduced the purposes and goals of the standard-setting event and

the methods, roles, and responsibilities of individuals involved in the event. The small group training, followed by the large group training, was given by room facilitators. In the small group training, the room facilitators emphasized the tasks to be performed and answered any questions that the panelists had following the training. Room facilitators also answered standard-setting related questions generated from panelists at their tables; however, policy-related questions were directed to KSDE staff.

Table VI-1 shows the demographic composite of panels by grade. Despite the efforts put into recruiting a diversified panelist group, the results are not as diverse as desired. In total, 40 panelists participated in the standard-setting event.

Table VI-1. Summary of Science Panelists' Demographic Information

		Grade		
		5	8	11
Demographics	Categories	% (n = 13)	% (n = 11)	% (n = 15)
Gender	Male	15.38	9.09	40.00
	Female	84.61	90.91	60.00
Race/ethnicity	Native American	0.00	0.00	0.00
	Asian/Pacific Islander	0.00	0.00	0.00
	Black	0.00	0.00	0.00
	Hispanic or Latino	9.09	0.00	0.00
	White	84.61	90.91	93.33
	Other	9.09	9.09	6.67
Teaching experience	1–3 years	0.00	0.00	0.00
	4–6 years	46.15	18.18	0.00
	7–12 years	23.08	18.18	33.33
	>12 years	30.77	63.64	66.67
Current assignment	Classroom teacher	100.00	100.00	73.33
	Educator (non-teacher)	0.00	0.00	13.33
	Other	0.00	0.00	13.33
Work setting	Urban	30.77	18.18	20.00
	Suburban	30.77	45.45	20.00
	Rural	38.46	36.36	60.00

VI.2.5.2. Completing the science exam. To provide a frame of reference for considering student performances, the panelists took the science test in a shorter timeframe than was used operationally. The panelists took the test in the KITE system using Chromebooks supplied by the

school. Students used Chromebooks during operational testing, so their use by the panelists mirrored the test-taking procedures followed by students. Panelists used the log-in information from the facilitator to log in to the KITE system, just as a student did for his or her operational test. The panelists were given 45 minutes to finish the test and were encouraged to think about how students might have experienced these items. After getting a feel for item and test difficulty through taking the test, the panelists discussed the item and test difficulty.

VI.2.5.3. "Just-barely" student activity and discussion. The just-barely student activity defines the performances of students who just barely reach Level 2, Level 3, and Level 4, as defined by the PLDs. The purpose of this activity is for panelists to focus on and develop a common understanding of just-barely Level 2, Level 3, and Level 4 knowledge and skills. The PLDs represent a wide range of content knowledge and skills for all students within an achievement level. The just-barely activity pinpoints the knowledge and skills of the students at the very bottom of that range—those whose scores would put them just barely in the level. The student score at the bottom of the level defies that cut score.

Panelists were guided to use the just-barely worksheets to help them defining the performances of students in this area. They used the just-barely worksheets to answer this question: "What knowledge and skills does a this student have that a student who is at the top of the lower adjacent level not have?" They started working individually and then had a group discussion. A list of just-barely performances for each achievement level was approved by panelists at the end of this activity.

VI.2.5.4. Bookmark practice. The purpose of this practice is to let panelists familiarize themselves with the Bookmark procedures. Using a practice OIB of 10 items, the practice item-map table, and the practice item-dotplot sheet, the panelists reviewed the practice items and considered the following questions.

- What do students have to know and be able to do to answer this item correctly?
- What makes this item more difficult than the ones preceding it?

Panelists discussed the knowledge and skills required to correctly answer the first item on the practice OIB using the questions listed above. They were guided to refer to their just-barely student list for Level 3 and ask themselves if two-thirds of the just-barely achieving Level 3 students would be able to answer this item correctly. For items that have more than one score point, they asked themselves, "Would two-thirds of just-barely Level 3 students be able to get this score point or higher?" If they mostly agreed that the answer is yes, they moved on to the second item. Panelists did this until reaching an item for which they said no. They placed the bookmark on that item. They repeated the same procedures to place the Level 4 and Level 2 bookmarks.

After panelists were comfortable with the Bookmark rating procedure, they moved to the actual rating. The actual rating has three rounds. After the first round, the group minimum, maximum, and median bookmark values are calculated. Based on this information, the panelists in one group (i.e., at one table) will discuss the differences and similarities within the group. The

panelists do not need to agree with everyone but do need to listen with an open mind. During the process, it is expected that the range in bookmarks will converge over rounds, but it is not required that all panelists come to consensus. After this discussion, the panelists place their Round 2 bookmarks. After Round 2, the psychometricians calculate the same information they did after Round 1. This time, all the tables will review the results together and have another discussion to identify any areas of disagreement. The impact data, data that show the distribution of students in each level, are also provided for panelists to consider. During Round 3, panelists have a third opportunity to change their bookmarks and make a final recommendation. Detailed bookmarking activities will be discussed in sections VI.2.5.6 through VI.2.5.8.

VI.2.5.5. Identify operational item knowledge and skills. Using the actual OIB, panelists reviewed each item and made notes in their OIBs to identify operational item knowledge and skills. They answered these questions for each item.

- What do students have to know and be able to do to answer this item correctly?
- What makes this item more difficult than the ones preceding it?

The panelists were reminded to consider both the PLDs and the just-barely descriptions. They could also refer to the KCCRS for science. The goal was to outline the knowledge and skills required to answer the items.

VI.2.5.6. Setting cut score: Round 1. Panelists checked out item map tables, item dot plots, and OIBs. They filled out the readiness form (Appendix L) before the Round 1 rating. All panelists responded yes to all the questions on the form before bookmark placing began. Many of the items in the OIB were ones panelists had seen from the example test; others were new items. They started with item 1 and kept going until they felt they had reached a point where two-thirds of the just-barely Level 3 students would not be able to answer the item correctly. The panelists put a bookmark on that page and also recorded the page number on the bookmark placement form. Once they placed their bookmarks, the panelists were reminded to look at a few items that came after the marked item to be sure they had their bookmarks in the right place. They were also reminded to consider the KCCRS for science, PLDs, the just-barely lists, and their knowledge and skills notes. After placing the Level 3 bookmark, panelists continued to place Level 4 then Level 2 bookmarks where they felt that two-thirds of just-barely Level 4 and Level 2 students would not be able to answer the item correctly.

Three rules to follow regarding panelists' bookmark placements are as follows:

1. If a just-barely Level 2 student would answer an item correctly, then a just-barely Level 3 student would also answer that item correctly. If a just-barely Level 3 student would answer an item correctly, then a just-barely Level 4 student would also answer that item correctly.
2. If a just-barely Level 4 student would not answer an item correctly, then a just-barely Level 3 student would not answer that item correctly either. If a just-barely Level 3 student would not answer an item correctly, then a just-barely Level 4 student would not answer that item correctly either.
3. Items are ordered by difficulty from easiest to hardest in the booklet. The Level 2

bookmark page should be the earliest, and the Level 4 bookmark page should be the latest among the three level bookmark pages.

Panelists completed this work independently. Table VI-2 presents the medians of Rounds 1 through 3 bookmark placements among panelists for grades 5, 8, and 11. After completing their bookmark placements, panelists submitted their bookmark placements to their facilitator and completed the evaluation form Parts I–IV (Appendix M).

VI.2.5.7. Setting cut score: Round 2. Panelists for each grade level first reviewed and discussed Round 1 bookmark results. Panelists were asked to consider questions like these:

- How tough or easy are you as a panelist?
- Were you stricter or more lenient than your tablemates?
- Were you consistently strict or lenient across all three bookmark placements, or did you vary?
- How consistent were panelists at your table?

Then they used some time to think about how their individual ratings compared to others. The guiding questions included the following:

- Why did you place your bookmark where you did?
- Where is the best place to separate the knowledge and skills of students at the just-barely level and above from students who are just below there?

Panelists were encouraged to use information from Round 1 results to inform themselves and to give themselves opportunities to reconsider ratings. They were instructed to consider the KCCRS for science, PLDs, the just-barely attributes of students, and their item knowledge and skill notes when placing bookmarks. The same procedures of placing bookmarks as in Round 1 were taken. Panelists were clearly told that they could change their bookmark placements or that they could keep their bookmark placements the same. Medians of Round 2 cut scores are presented in Table VI-2. After completion of the bookmark placement, panelists submitted their bookmark placement forms to their facilitators and completed the evaluation form Part V.

VI.2.5.8. Setting cut score: Round 3. Panelists reviewed and discussed bookmark placements from Round 2 to consider the best place to separate the knowledge and skills of students at the just-barely level and above from students who are just below there, taking the following questions into consideration:

- What is the range of the bookmark placements?
- How did the range for Round 2 change compared to Round 1?
- How does your bookmark placement compare to the room average placement?

Another piece of information for panelists to consider was the impact data based on Round 2. With information provided by the impact data, panelists got an idea about the percentage of student scores that would be classified in each achievement level (Level 1, Level 2, Level 3, and Level 4), given the bookmarks that came out of Round 2. These percentages were for students who actually took the 2017 Kansas assessment. Facilitators showed panelists the outcomes, given their current recommendations. Finally, a graph showing how Kansas students fared on the NAEP science assessments in 2015 on the nearest grade level was presented to panelists to

provide an additional point of reference about the achievement of Kansas students in science. Grade 5 panelists saw grade 4 NAEP; grades 8 and 11 panelists saw grade 8 NAEP.

Considering all the data presented as well the PLDs and just-barely attributes of students, panelists were guided to think about the following questions before placing bookmarks for the last round:

- If you believe your placement was too lenient or too strict compared to others, what could you do differently?
- Were all three of your bookmarks higher or lower than the median? That is, were you consistently lower? Or perhaps you were lower on one bookmark placement but higher on another bookmark placement? What does this tell you?
- Additionally, after thinking about the impact data, how does the percentage distribution match your experience with students?
- What will the results be if you stay with your current recommendations?

Panelists put all the information together and placed their best and final bookmarks, a recommendation of the final level cuts to the State Board. The medians of Round 3 cuts are presented in Table VI-2. Panelists submitted their bookmark placement and completed the remaining sections of the evaluation form. Table VI-3 shows a summary of panelists' responses to questions regarding the results of cut scores in the evaluation form. These evaluation questions were abbreviated and modified for presentation purposes. Refer to Appendix M for the actual evaluation questions.

Table VI-2. Rounds 1–3 Medians of Bookmark Placements by Grade

	Grade 5 level			Grade 8 level			Grade 11 level		
Round	2	3	4	2	3	4	2	3	4
1	11	23	40.5	10	28.0	44.5	10	20	44
2	10	22	40.0	10	28.5	46.5	10	26	42
2	9	22	37.0	14	28.0	47.0	10	25	40

Table VI-3. Summary of Panelists' Perceptions about Cut-Score Results in Evaluation Survey

Questions	Means		
Grade	5	8	11
Strongly Disagree (1) to Strongly Agree (6)			
Grade-level group results for the Level 2 cut scores			
Impact result for Level 2 is reasonable.	5.2	4.5	4.9
Cut score for Level 2 is appropriate.	5.2	4.9	5.2
Cut score for Level 2 is defensible.	5.2	5.0	5.0
Grade-level group results for the Level 3 cut scores			
Impact result for Level 3 is reasonable.	5.3	5.2	4.6
Cut score for Level 3 is appropriate.	5.3	5.2	4.6
Cut score for Level 3 is defensible.	5.4	5.3	4.9
Grade-level group results for the Level 4 cut scores			
Impact result for Level 4 is reasonable.	5.3	5.0	4.9
Cut score for Level 4 is appropriate.	5.3	5.2	4.9
Cut score for Level 4 is defensible.	5.3	4.9	5.1

VI.2.5.9. Articulation training and articulation. Articulation training helps panelists understand and become familiar with the articulation process. The articulation leader provided the training for articulation covering the following topics:

- Articulation purpose;
- Panelist roles and responsibilities;
- Panelist expectation on the level cut consistency across grades;
- Articulation procedure;
- Standard error of judgment (SEJ); and
- Reasonable level cut-score adjusting range.

During the articulation, three articulation panelists from grade 11, three from grade 8, and two from grade 5 worked together to articulate the level cut scores across grades. First, the articulation leader presented the Round 3 level cuts, the impact data of all grades, and the reasonable ranges that the level cuts could be adjusted within. Then, panelists discussed these results. Next, the panelists worked as a group to adjust the level cut scores by looking at the effect of changing level cut scores on the impact data. Finally, panelists completed the evaluation form of the articulation section (see Appendix N).

Articulation discussion was guided using the following questions:

- What are the differences between the impact data and your expectation on these cuts?
- Why do you think the impact data do not match your expectations?

After the discussion, the articulation leader showed the min, max, cuts, and SEJ of the Round 3 results of all grades and introduced the reasonable range for which the panelists could adjust cut scores within. The reasonable range of adjusting cuts is the cut scores ±1 SEJ.

The articulation leader answered any questions that panelists had regarding the articulation before they started the articulation discussion as a group. Then, the articulation leader led the discussion by asking the panelists how they would adjust the level cut scores to meet their expectation. After the discussion, the articulation leader used a data tool that allowed panelists to see the change of the impact data after adjusting the level cut score. Using this tool, articulation panelists were led to get one set of level cut scores of all grades agreed on by all panelists. In this process, the articulation leader reminded the panelists to adjust their cut scores within the reasonable ranges.

Table VI-4 presents the articulation data, showing the minimum, median, and maximum cut thetas at each performance level for each grade.

Table VI-4. Articulation Data by Grade and Level

	Grade 5 level			Grade 8 level			Grade 11 level		
	2	3	4	2	3	4	2	3	4
Min	−1.00	−0.10	1.10	−0.75	0.30	1.20	−0.70	0.20	1.15
Median	−0.94	−0.03	1.16	−0.60	0.40	1.51	−0.55	0.32	1.45
Max	−0.90	0.00	1.30	−0.36	0.53	1.52	−0.47	0.34	1.60

VI.3. Challenging and Aligned Academic Achievement Standards

The KAP grade-level academic achievement standards are drafted to align with the state content standards: the KCCRS for ELA, mathematics, science, and HGSS. CETE content experts worked alongside KSDE staff to define the PLDs. The iterative process ended when both sides agreed that the expected performances adhered to KCCRS content standards, as well as to cognitive demands, and that the overall expectation properly reflected the rigor of the KCCRS. Then, the PLDs were presented to Kansas educators for review and approval. As described in the science standard-setting section, PLDs are the basis of the cut scores.

VI.4. Reporting

For each tested subject, the KAP assessment provides separate score reports to students, schools, and districts (see Appendix O). The content of these reports includes overall performance and performance by content standards. These statistics are presented using various graphs, colors, and symbols so they are easy to read. To assist readers in interpreting the information in the reports, descriptions of what students should be able to do in each subject area are presented with the statistics. As stated by Petersen, Kolen, and Hoover (1989), providing score interpretations in score reports can minimize misinterpretations and unwarranted inferences. It is as critical to help readers understand the meaning of the statistics as it is to report the values.

Although these reports are intended for different groups (e.g., students, schools, and districts), the content of these reports is uniform. Presentation and text are adjusted according to group, but the symbols and interpretation of those symbols are consistent across reports. This design eases

educators' reviewing burden and helps them explain score reports to parents.

VI.4.1. Group masking. When group *n* counts are very small, individual students may be identified through demographic information, even on roll-up summary reports. Various types of suppression logic are used to protect individual identities. One way is to report student score results by percentage ranges instead of the actual observed 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%). The other way to report is to suppress reports when the number of students on the group is small.

VI.4.2. Student reports. A sample of an ELA student report is presented in Appendix O. In the report, a student's performance level is placed immediately after student identifiers so that it is the first information presented. Next are the student's scaled score and comparisons with students in the same school, district, and state (i.e., the score meters), as well as a brief summary of the PLDs that describe what this student should be able to do. Score meters report the medians of school, district, and state performances. The median is used because it is more robust to outliers than the mean in describing the central tendency of a group.

A student's overall-score performance level represents a student's performance on all sections of the test. The final section of the report is an overall policy-level description for each performance level.

The first section of the second page reports the student's performance by subscores. This information indicates strengths and weaknesses on different claims or targets. Each claim/target represents a group of test items that assess related skills. Some items of a test are counted in multiple categories. Subscore information is not available for science or HGSS reports.

The bottom of the second page shows the *SE* of scaled scores and *SE*s of school, district, and state median scores. The *SE* reported on student scores is the CSEM derived from the IRT scaling model. It indicates how much a student score might vary if the student took many equivalent versions of the test. The *SE*s of group scores (school, district, and state) account for sampling error but not for measurement error.

The *SE* of the median is computed using equation VI-1. It is equivalent to the *SE* of the mean but multiplied by an extension factor of 1.253 to account for the additional sampling variability of the median.

$$SE_{median(x)} = 1.253 * \frac{S_{\bar{x}}}{\sqrt{N}}, \qquad \text{(VI-1)}$$

where $SE_{median(x)}$ is the *SE* of the median of the group scores, $S_{\bar{x}}$ is the *SD* of the group's observed scores, and *N* is the number of students in the group.

Timeline for delivering student reports. The KAP testing window ended on April 28, 2017. In May 2017, student reports were available for all students who took the KAP ELA and mathematics tests. The standard-setting event and the required approval of cut scores delayed the delivery of science student reports to educators from summer to early fall 2017.

VI.4.3. School and district reports. While student reports focus on individual student performance, school and district reports focus on group-level performances. Information provided in the school and district reports aggregates student performances at the given level (see Appendix O).

School reports provide summary information of the same subject by grades. On the first page, bar graphs show a school's median scaled scores of three grades, along with scores of the school's district and state overall performances. District and state median scaled scores are reference for schools to interpret their standings. The *SE*s are given at the bottom of the first page. The second page shows the percentage of students in each of the four performance levels; again, district and state results are provided for reference. The bar graphs use four different colors to represent the different performance levels, allowing readers to distinguish performance-level outcomes instantly. The next section of the school report presents the school's performance by claim/target: student performances by content standards and a summary of students' relative strengths and weaknesses in the different content standards.

District reports use the same layout and provide the same information as school reports; however, only state data are provided as the reference group.

VI.4.4. Interpretive guides. Besides adding descriptions to score reports, two score interpretive guides, *2017 Educator Guide – Understanding the Kansas Assessment Program Score Report* and *2017 Parent Guide – Understanding the Kansas Assessment Program Score Report*, are available for educators and parents to download from the KAP website.

VI.4.5. Letters from the Commissioner of Education. The letters to Kansas educators and parents from Dr. Randy Watson, Kansas Commissioner of Education, are an important part of the interpretive guides. Copies of these two letters are provided in Appendix P.

VII. References

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison Wesley.

Brennan, R. L. (2004). *Manual for BB-CLASS: A computer program that uses the Beta-binomial model for classification consistency and accuracy, Version 1.1 (Center for Advanced Studies in Measurement and Assessment Research Report No. 9)*. Iowa City: University of Iowa.

Cai, L. (2013). *flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications, Inc.

Common Core State Standards Initiative. (2010). *Common core state standards for English language arts and literacy in history/social studies, science, and technical subjects.* Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart and Winston, Inc.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement*, pp. 443–507.

Forte, E. (2013, June). *Next generation alignment approaches: Needs and promising directions.* Paper presented at the National Conference on Student Assessment, Baltimore, MD.

Forte, E. (2016). *Evaluating alignment in large-scale standards-based assessment systems*. Washington, DC: Technical Issues in Large Scale Assessment SCASS of CCSSO.

Forte, E., Nebelsick-Gullett, L., Deters, L., Buchanan, E., Herrera, B., Morris, J., & Phlegar, J. (2016). *Kansas Assessment Program alignment evaluation report 2015–2016*. edCount LLC.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347–360.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 47–76). New York, NY: Routledge.

Houts, C. R., & Cai, L. (2013). *flexMIRT user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring*. Chapel Hill, NC: Vector Psychometric Group.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rate using an effect size

measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349.

Kansas State Department of Education. (2010a). *Kansas college and career ready standards: English language arts and literacy in history/social studies, science, and technical subjects.* Retrieved from http://community.ksde.org/LinkClick.aspx?fileticket=tzz1aDOC0v8%3D&tabid=5559&mid=13575.

Kansas State Department of Education. (2010b). *Kansas college and career ready standards: Mathematics grades K–12 with Kansas 15%.* Retrieved from http://community.ksde.org/LinkClick.aspx?fileticket=iX3kWvXtgRY%3d&tabid=5276&mid=13067.

Kansas State Department of Education. (2013). *Kansas standards for history, government, and social studies*. Retrieved from http://www.ksde.org/LinkClick.aspx?fileticket=zNGRyc6vESw%3D&tabid=472&portalid=0&mid=1587.

Kansas State Department of Education. (2017). *Kansas assessment examiner's manual 2016–2017*. Retrieved from http://ksassessments.org/sites/default/files/documents/Kansas_Assessment_Examiners_Manual.pdf.

Kansas State Department of Education. (2015). *Tools and accommodations for the Kansas assessment program (KAP)*. 2015–2016. Website updated to 2016–2017 manual.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Koretz, D. (2008). *Measuring up: What educational testing really tells us* (pp. 7–14). Cambridge, MA: Harvard University Press.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179–197.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 249–282). Mahwah, NJ: Lawrence Erlbaum.

National Research Council of the National Academies. (2012). *A framework for K–12 science education: Practice, crosscutting concepts, and core ideas.* Washington, DC: The National Academies Press. Retrieved from https://www.nap.edu/read/13165/chapter/1.

Next Generation Science Standards [NGSS]. (2013). *The next generation science standards: Executive summary.* Retrieved from NGSS https://www.nextgenscience.org/sites/default/files/Final%20Release%20NGSS%20Front%20Matter%20-%206.17.13%20Update_0.pdf.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York, NY: American Council on Education/Macmillan.

Pitoniak, M. J., & Morgan, D. L. (2012). Setting and validating cut scores for tests. In C. Secolsky & D. B. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (pp. 343–366). New York, NY: Routledge.

Press, W. H., Flannery B. P., Teukolsky S. A., & Vetterling, W. T. (1989). *Numerical recipes*. New York, NY: Cambridge University Press.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.

School performance accreditation system; curriculum standards; student assessments; school site councils, Kan. Stat. Ann. §72-6479 (2015). Retrieved from http://kslegislature.org/li/b2015_16/statute/072_000_0000_chapter/072_064_0000_article/072_064_0079_section/072_064_0079_k/.

U.S. Department of Education (2015). Peer Review of State Assessment Systems: Non-regulatory guidance for states for meeting requirements of the Elementary and Secondary Education Act of 1965, as amended. Retrieved from https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Council of Chief State School Officers [CCSSO] and National Institute for Science Education Research Monograph No. 6). Washington, DC: CCSSO.

VIII. Appendix A: Test Administration and Security Training

Tickets and Security

- New process for "tickets"
- Student uses his/her designated username and password for the entire testing season
- Daily access codes will be posted in the Educator Portal
- Plan for your secure transmission of this information each day
- Details in the Examiner's Manual

ETHICS

Test Item Security

Teacher/test proctor **may not store or save on computers or personal storage devices** any test items; test items may not be shared via email or other file sharing systems or reproduced by any means.

Still happening, so please reinforce!

District Test Coordinator Responsibilities

- List of responsibilities included with fact sheet
- Review carefully to be certain all expectations have been met

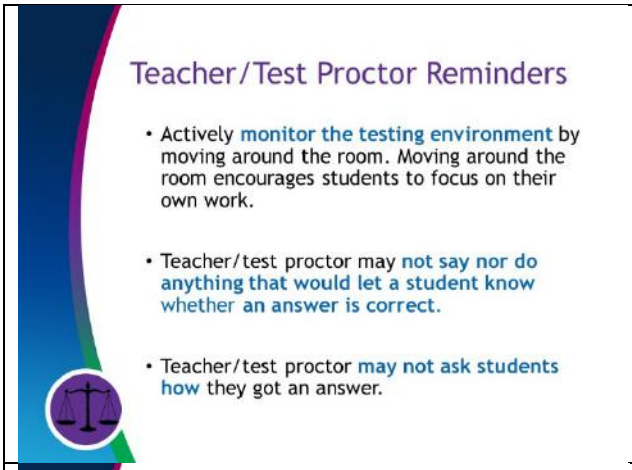
Reporting Item Issues

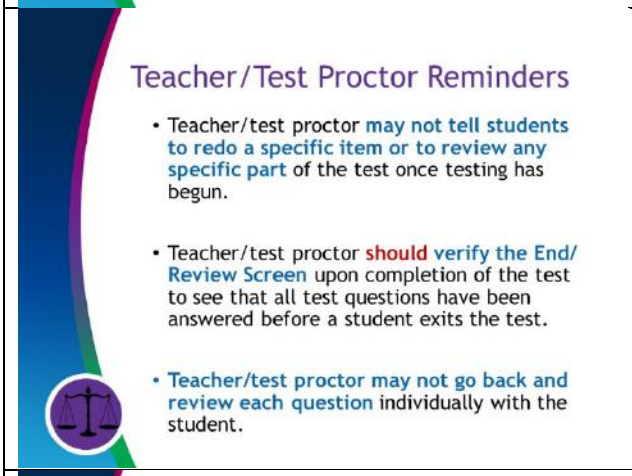
Problems reported to District Test Coordinator who contacts KSDE

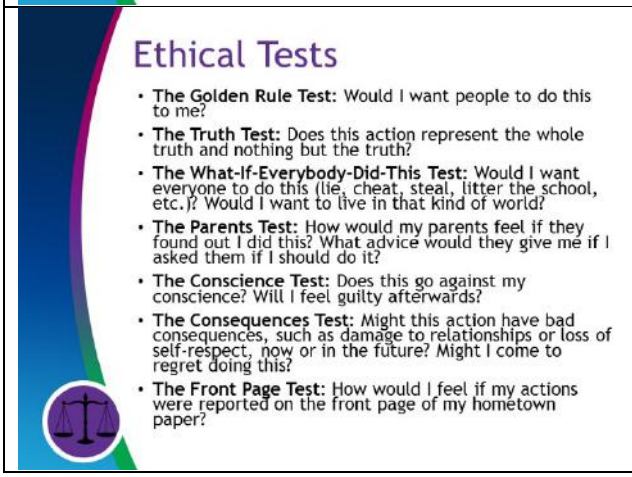
Process outlined on the Test Security Fact Sheet

Teacher/Test Proctor Responsibilities

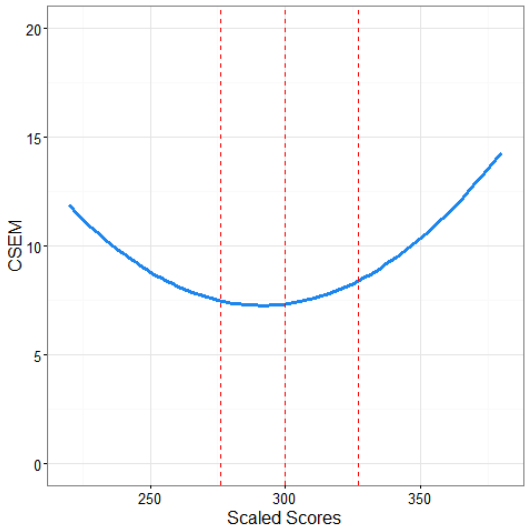
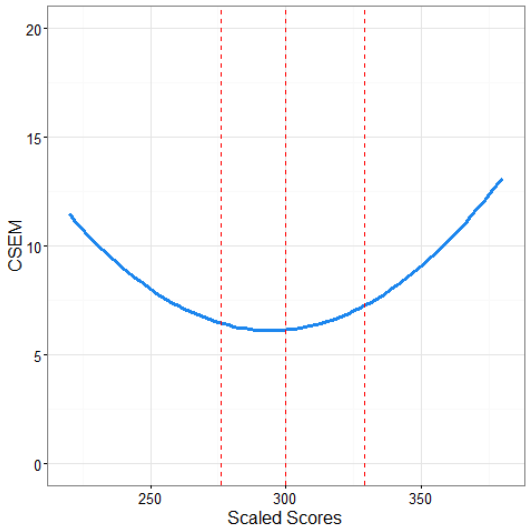
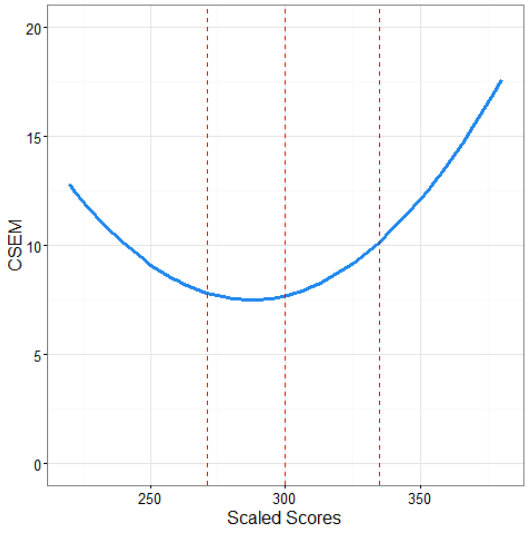
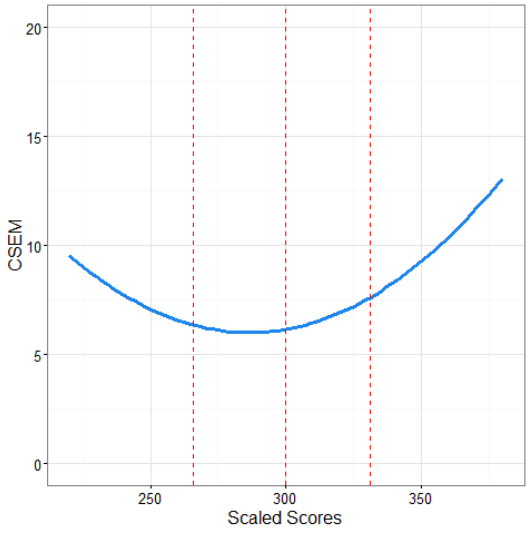
- List of responsibilities included with fact sheet
- Must also review the Examiner's Manual and abide by all directions
- Fact sheet also includes list for Building Test Coordinators

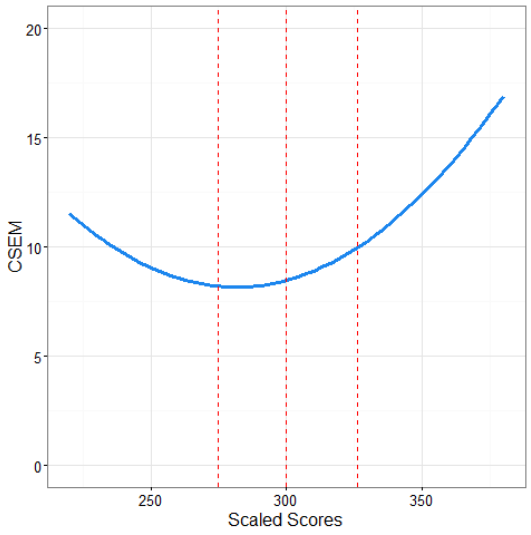
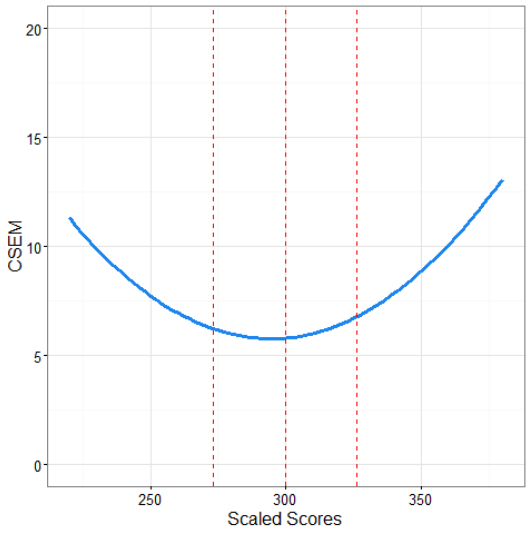
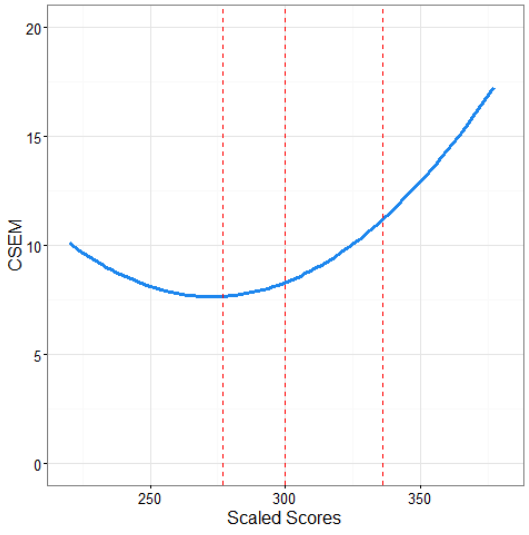
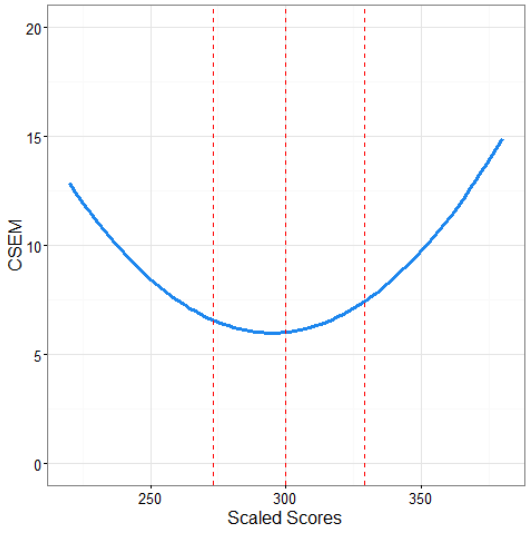

Teacher/Test Proctor Reminders
• Actively monitor the testing environment by moving around the room. Moving around the room encourages students to focus on their own work.
• Teacher/test proctor may not say nor do anything that would let a student know whether an answer is correct.
• Teacher/test proctor may not ask students how they got an answer.

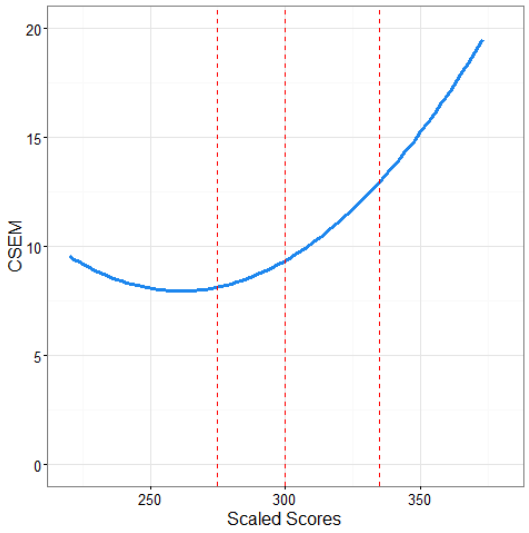
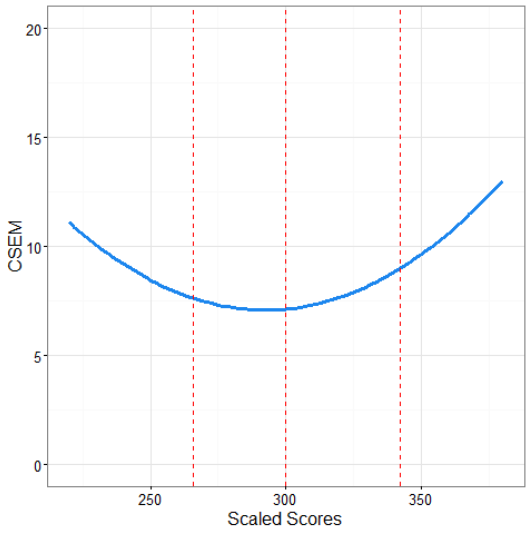
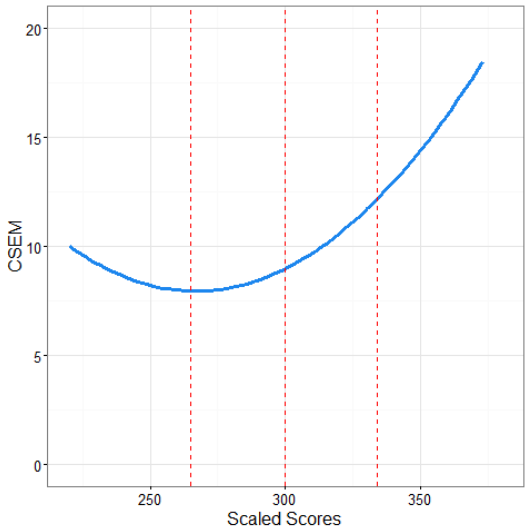
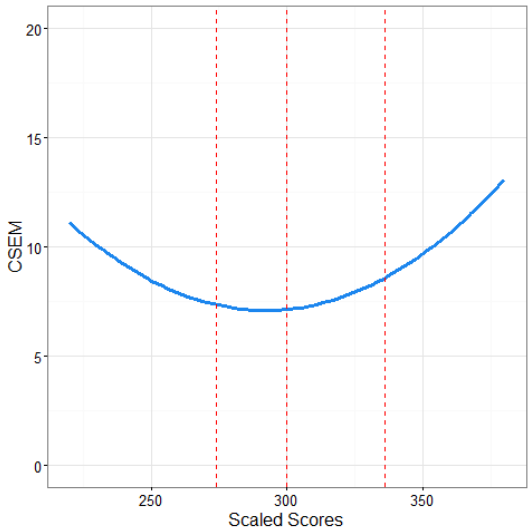

Teacher/Test Proctor Reminders
• Teacher/test proctor may not tell students to redo a specific item or to review any specific part of the test once testing has begun.
• Teacher/test proctor should verify the End/ Review Screen upon completion of the test to see that all test questions have been answered before a student exits the test.
• Teacher/test proctor may not go back and review each question individually with the student.

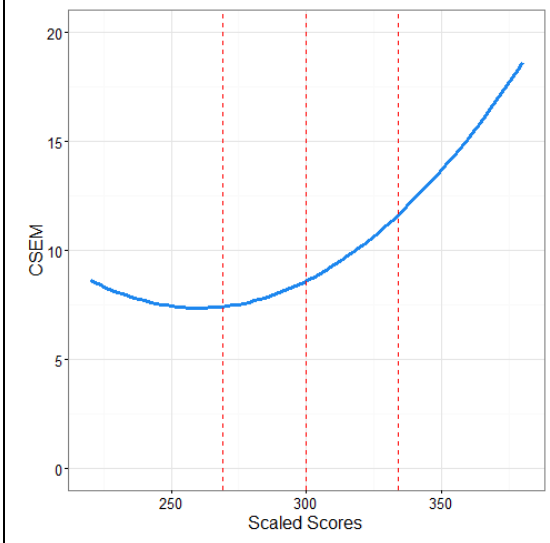
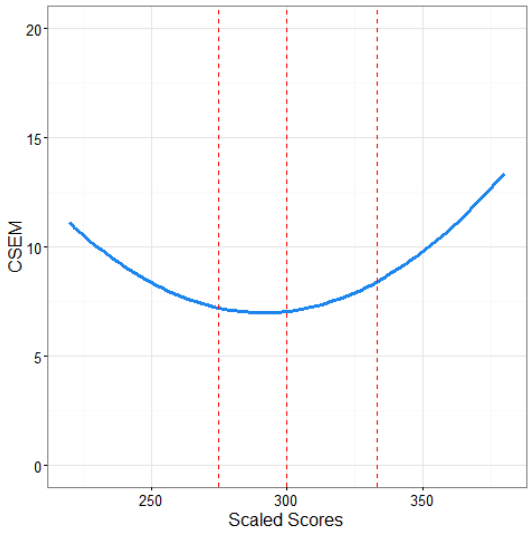

Ethical Tests
• The Golden Rule Test: Would I want people to do this to me?
• The Truth Test: Does this action represent the whole truth and nothing but the truth?
• The What-If-Everybody-Did-This Test: Would I want everyone to do this (lie, cheat, steal, litter the school, etc.)? Would I want to live in that kind of world?
• The Parents Test: How would my parents feel if they found out I did this? What advice would they give me if I asked them if I should do it?
• The Conscience Test: Does this go against my conscience? Will I feel guilty afterwards?
• The Consequences Test: Might this action have bad consequences, such as damage to relationships or loss of self-respect, now or in the future? Might I come to regret doing this?
• The Front Page Test: How would I feel if my actions were reported on the front page of my hometown paper?

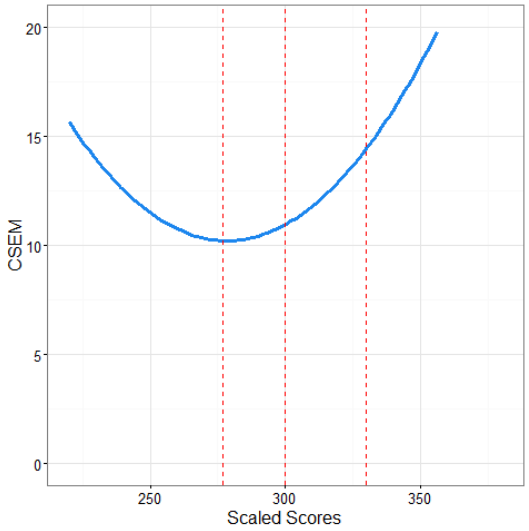
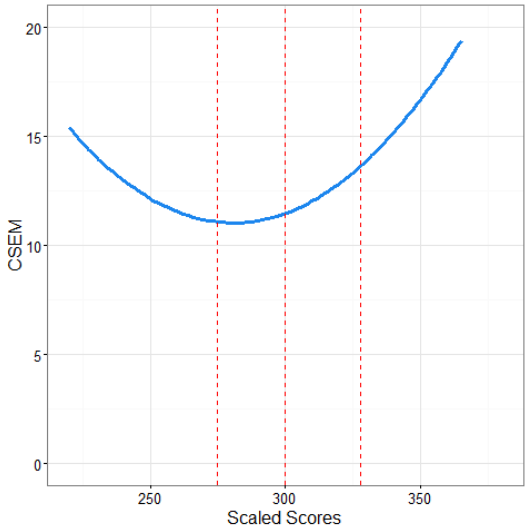
IX. Appendix B: Conditional Standard Error of Measurement (CSEM)

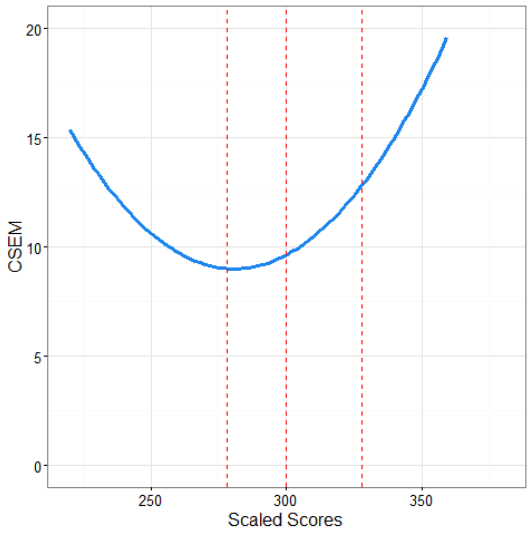
Grade	ELA CSEM	Math CSEM
3		
4		

5	 CSEM Scaled Scores	 CSEM Scaled Scores
6	 CSEM Scaled Scores	 CSEM Scaled Scores

7	 CSEM Scaled Scores	 CSEM Scaled Scores
8	 CSEM Scaled Scores	 CSEM Scaled Scores

10	 20 15 10 5 0 CSEM 250 300 350 Scaled Scores	 20 15 10 5 0 CSEM 250 300 350 Scaled Scores

Grade	Science CSEM
5	
8	

11	 20 15 10 5 0 CSEM 250 300 350 Scaled Scores

X. Appendix C: Frequency Distribution and CSEM of Scale Scores

For all tables in Appendix C, the inner-table horizontal lines separate scale scores falling in different performance levels. To be specific, the first line separates performance Levels 1 and 2; the second line separates performance Levels 2 and 3; and the third line separates performance Levels 3 and 4.

Table C-1. Grade 3 Scale-Score Distribution for ELA

Grade	Scale score	CSEM	*N*
3	220	13.3	32
3	222	12.0	46
3	228	11.0	53
3	233	10.2	126
3	237	9.6	180
3	241	9.1	301
3	244	8.8	381
3	248	8.5	478
3	251	8.2	587
3	254	8.0	724
3	257	7.8	755
3	259	7.7	818
3	262	7.6	857
3	265	7.5	932
3	267	7.4	926
3	270	7.3	956
3	272	7.3	990
3	274	7.3	1,037
3	277	7.3	1,074
3	279	7.2	1,032
3	280	7.9	38
3	281	7.2	1,123
3	283	7.8	64
3	284	7.2	1,066
3	285	7.6	97
3	286	7.2	1,033
3	287	7.5	82
3	288	7.3	1,085
3	290	7.3	729
3	291	7.3	378
3	292	7.4	348
3	293	7.3	828
3	295	7.4	1,153
3	297	7.3	600
3	298	7.4	560
3	299	7.3	689
3	300	7.5	395

Grade	Scale score	CSEM	*N*
3	301	7.3	432
3	302	7.3	360
3	303	7.6	285
3	304	7.3	879
3	305	7.8	159
3	306	7.4	892
3	308	7.5	611
3	309	7.4	408
3	310	8.1	15
3	311	7.5	962
3	313	7.6	930
3	314	8.3	4
3	316	7.6	879
3	318	7.8	889
3	321	7.9	919
3	323	8.0	465
3	324	8.0	396
3	326	8.2	777
3	329	8.3	766
3	332	8.5	710
3	335	8.7	733
3	338	9.0	325
3	339	9.0	310
3	342	9.3	560
3	345	9.6	277
3	346	9.6	233
3	349	9.9	270
3	350	10.0	174
3	354	10.4	350
3	358	10.9	262
3	363	11.6	110
3	364	11.6	109
3	369	12.4	142
3	376	13.7	100
3	380	15.3	100

Table C-2. Grade 4 Scale-Score Distribution for ELA

Grade	Scale score	CSEM	*N*
4	220	14.3	25
4	221	12.9	21
4	227	11.8	39
4	232	11.0	54
4	237	10.3	91
4	241	9.7	150
4	244	9.3	36
4	245	9.3	164
4	247	9.8	2
4	248	8.9	302
4	251	8.6	385
4	254	8.3	457
4	255	9.0	12
4	257	8.1	520
4	258	8.7	31
4	260	7.9	551
4	261	8.4	64
4	262	7.8	532
4	264	8.2	123
4	265	7.7	495
4	267	7.7	760
4	269	7.9	102
4	270	7.6	761
4	272	7.6	963
4	275	7.6	988
4	277	7.6	1,077
4	279	7.6	1,095
4	282	7.5	1,098
4	284	7.5	1,148
4	286	7.5	1,120
4	289	7.5	1,192
4	291	7.5	1,131
4	293	7.6	1,119
4	294	8.0	37
4	296	7.6	1,219
4	298	7.7	1,160
4	299	8.3	6
4	300	7.7	1,223
4	302	8.5	1
4	303	7.8	1,259
4	305	7.9	1,176
4	307	8.0	532
4	308	8.0	677
4	310	8.2	1,237

Grade	Scale score	CSEM	*N*
4	313	8.3	1,274
4	315	8.5	1,203
4	318	8.7	1,204
4	321	8.9	1,189
4	324	9.1	1,184
4	327	9.4	1,081
4	330	9.7	951
4	334	10.0	937
4	337	10.4	430
4	338	10.4	476
4	341	10.9	334
4	342	10.9	426
4	346	11.4	645
4	350	12.0	294
4	351	12.1	296
4	355	12.8	224
4	356	12.8	254
4	361	13.7	144
4	362	13.7	169
4	368	14.9	242
4	376	16.4	140
4	380	18.5	201

Table C-3. Grade 5 Scale-Score Distribution for ELA

Grade	Scale score	CSEM	*N*
5	220	12.4	42
5	225	11.5	38
5	230	10.9	83
5	234	10.4	38
5	235	10.3	102
5	238	9.8	229
5	242	9.4	358
5	245	9.1	431
5	248	8.8	160
5	249	8.8	383
5	252	8.5	613
5	254	8.3	195
5	255	8.3	509
5	257	8.1	749
5	260	7.9	772
5	263	7.8	809
5	265	7.7	887
5	268	7.6	880
5	270	7.6	836
5	271	9.2	3
5	273	7.5	903
5	274	9.0	18
5	275	7.5	940
5	277	8.9	32
5	278	7.6	947
5	279	8.8	32
5	280	7.6	976
5	282	7.9	741
5	283	7.7	385
5	285	7.9	1,137
5	287	8.1	644
5	288	8.2	522
5	290	8.2	1,226
5	293	8.4	1,230
5	295	8.5	985
5	296	8.5	272
5	298	8.6	1,072
5	299	8.7	226
5	300	8.5	442
5	301	8.7	600
5	302	8.9	170
5	303	8.5	873
5	304	9.2	153
5	305	9.1	122

Grade	Scale score	CSEM	*N*
5	306	8.6	996
5	308	8.7	1,256
5	311	8.7	1,121
5	314	8.8	1,039
5	315	10.0	22
5	316	8.9	574
5	317	8.8	480
5	318	10.4	9
5	319	9.0	1,013
5	322	9.2	1,026
5	323	10.7	5
5	325	9.4	935
5	327	11.2	4
5	328	9.6	897
5	331	11.7	2
5	332	9.8	866
5	335	10.1	783
5	336	12.2	2
5	339	10.4	741
5	343	10.8	664
5	347	11.3	609
5	352	11.9	514
5	357	12.5	421
5	362	13.4	286
5	369	14.5	230
5	376	16.0	50
5	377	15.9	75
5	380	18.1	145

Table C-4. Grade 6 Scale-Score Distribution for ELA

Grade	Scale score	CSEM	*N*
6	220	10.2	110
6	221	9.7	56
6	225	9.2	101
6	228	9.0	31
6	229	8.9	115
6	232	8.6	202
6	234	10.6	1
6	235	8.3	281
6	238	8.1	343
6	239	9.8	3
6	241	7.9	401
6	242	9.6	4
6	243	8.1	142
6	244	7.8	304
6	246	7.9	153
6	247	7.8	328
6	249	7.8	170
6	250	7.8	412
6	251	7.6	191
6	252	7.6	345
6	253	8.6	54
6	254	8.0	294
6	255	7.6	298
6	256	7.9	222
6	257	7.8	511
6	259	7.8	269
6	260	7.8	538
6	261	7.7	134
6	262	7.8	426
6	263	7.9	343
6	264	7.8	316
6	265	7.8	623
6	266	7.8	124
6	267	7.8	236
6	268	7.7	656
6	269	7.8	123
6	270	7.8	458
6	271	7.7	460
6	272	7.8	443
6	273	7.9	135
6	274	7.7	608
6	275	7.6	326
6	276	7.7	627
6	277	7.7	423

Grade	Scale score	CSEM	*N*
6	278	8.2	71
6	279	7.6	589
6	280	7.6	414
6	281	7.7	717
6	282	7.7	483
6	284	7.7	653
6	285	7.7	478
6	287	7.7	1,161
6	288	8.6	2
6	289	7.8	662
6	290	7.8	540
6	292	7.8	681
6	293	7.9	581
6	294	8.9	3
6	295	7.9	1,267
6	296	8.9	1
6	297	8.0	767
6	298	8.1	589
6	299	9.1	1
6	300	8.1	765
6	301	8.2	582
6	303	8.2	775
6	304	8.3	602
6	306	8.3	759
6	307	8.4	579
6	309	8.5	676
6	310	8.6	596
6	312	8.6	731
6	313	8.7	595
6	315	8.8	720
6	316	8.9	547
6	319	9.1	1,156
6	322	9.3	646
6	323	9.5	532
6	326	9.6	581
6	327	9.8	420
6	330	10.0	479
6	331	10.2	414
6	335	10.6	737
6	339	11.0	361
6	340	11.3	289
6	345	11.9	497
6	351	12.9	329
6	358	14.1	237
6	366	15.6	70
6	367	16.1	68
6	376	17.9	44

Grade	Scale score	CSEM	*N*
6	377	18.5	31
6	380	21.9	45

Table C-5. Grade 7 Scale-Score Distribution for ELA

Grade	Scale score	CSEM	*N*
7	220	10.7	135
7	224	10.1	95
7	228	9.6	129
7	232	9.2	196
7	235	8.9	280
7	238	8.6	344
7	241	8.3	517
7	244	8.1	581
7	247	7.9	642
7	250	7.8	701
7	252	7.7	773
7	254	7.9	3
7	255	7.6	834
7	257	7.8	7
7	258	7.5	892
7	259	7.7	4
7	260	7.5	911
7	262	7.7	6
7	263	7.5	947
7	265	7.5	978
7	268	7.5	1,029
7	270	7.6	1,063
7	273	7.7	1,140
7	275	7.8	1,078
7	276	8.0	108
7	278	7.9	1,212
7	280	8.1	1,008
7	281	8.2	256
7	283	8.3	914
7	284	8.4	389
7	286	8.5	902
7	287	8.5	483
7	289	8.7	728
7	290	8.7	643
7	292	9.0	627
7	293	8.9	800
7	295	9.3	471
7	296	9.1	899
7	299	9.4	1,331
7	302	10.1	237
7	303	9.5	1,177
7	306	9.8	1,264
7	310	10.0	1,276
7	313	10.2	1,225

Grade	Scale score	CSEM	*N*
7	314	11.7	47
7	317	10.5	1,158
7	319	12.3	29
7	321	10.8	1,048
7	324	13.1	13
7	325	11.2	1,046
7	330	11.6	916
7	335	12.1	797
7	337	15.1	6
7	340	12.7	609
7	344	16.3	4
7	345	13.4	546
7	351	14.3	453
7	352	17.8	1
7	358	15.4	355
7	365	16.8	231
7	373	18.4	168
7	380	20.4	207

Table C-6. Grade 8 Scale-Score Distribution for ELA

Grade	Scale score	CSEM	*N*
8	220	10.3	152
8	221	10.0	61
8	224	9.7	31
8	225	9.6	106
8	227	9.4	47
8	229	9.2	164
8	231	9.0	75
8	233	8.9	239
8	234	8.8	105
8	236	8.6	280
8	237	8.5	151
8	239	8.4	331
8	240	8.4	175
8	242	8.2	387
8	243	8.2	232
8	245	8.1	472
8	246	8.1	274
8	247	7.9	494
8	248	8.8	1
8	249	7.9	281
8	250	7.8	500
8	251	7.9	351
8	253	7.8	523
8	254	7.8	367
8	255	7.7	580
8	257	7.8	413
8	258	7.7	576
8	259	7.8	435
8	261	7.7	548
8	262	7.8	456
8	263	7.7	612
8	264	7.8	452
8	265	8.2	36
8	266	7.8	594
8	267	7.9	545
8	268	7.8	655
8	269	7.9	553
8	270	8.1	78
8	271	7.9	655
8	272	8.0	612
8	273	8.0	588
8	274	8.1	103
8	275	8.1	639
8	276	8.1	660

Grade	Scale score	CSEM	*N*
8	277	8.2	652
8	279	8.3	687
8	280	8.2	629
8	281	8.2	238
8	282	8.3	682
8	283	8.5	361
8	284	8.2	292
8	285	8.4	711
8	286	8.4	615
8	288	8.5	658
8	289	8.5	608
8	290	8.3	465
8	291	8.6	617
8	292	9.0	129
8	293	8.4	516
8	294	8.5	576
8	295	9.2	83
8	296	8.5	521
8	297	8.6	574
8	298	8.6	607
8	299	8.6	581
8	301	8.8	587
8	302	8.7	546
8	304	8.9	566
8	305	8.9	566
8	306	10.1	6
8	307	9.1	560
8	308	9.1	529
8	309	10.5	9
8	311	9.3	1,012
8	314	9.6	494
8	315	9.5	460
8	318	9.9	873
8	321	10.3	462
8	322	10.2	404
8	325	10.7	374
8	326	10.6	338
8	330	11.3	668
8	335	11.9	530
8	340	12.7	430
8	346	13.5	160
8	347	13.8	178
8	353	14.7	126
8	354	15.1	134
8	361	16.3	74
8	362	16.8	80
8	371	18.6	47

Grade	Scale score	CSEM	*N*
8	373	19.3	45
8	380	22.7	50

Table C-7. Grade 10 Scale-Score Distribution for ELA

Grade	Scale score	CSEM	*N*
10	220	9.2	203
10	221	8.8	109
10	225	8.5	161
10	228	8.2	219
10	231	8.0	290
10	234	7.8	341
10	237	7.7	436
10	240	7.5	482
10	242	7.4	577
10	245	7.3	609
10	247	7.3	626
10	249	7.2	644
10	252	7.2	727
10	254	7.2	739
10	255	7.5	14
10	256	7.1	731
10	258	7.4	6
10	259	7.1	783
10	260	7.4	6
10	261	7.1	803
10	263	7.2	855
10	265	7.2	414
10	266	7.2	476
10	267	7.4	26
10	268	7.2	887
10	270	7.3	951
10	272	7.4	620
10	273	7.4	444
10	275	7.4	1,135
10	277	7.5	1,103
10	279	7.6	513
10	280	7.6	644
10	282	7.8	1,234
10	284	7.9	352
10	285	7.9	868
10	287	8.0	1,002
10	288	8.0	168
10	289	8.1	474
10	290	8.1	737
10	292	8.2	467
10	293	8.2	706
10	295	8.4	1,077
10	296	8.5	100
10	298	8.5	1,041

Grade	Scale score	CSEM	*N*
10	299	8.7	17
10	300	8.8	27
10	301	8.7	1,032
10	302	9.0	11
10	303	9.0	15
10	304	8.9	1,087
10	305	9.2	5
10	306	9.3	6
10	307	9.1	500
10	308	9.1	597
10	309	9.5	2
10	310	9.3	436
10	311	9.4	564
10	313	9.9	1
10	314	9.6	496
10	315	9.7	556
10	317	9.9	472
10	319	10.1	500
10	321	10.3	418
10	323	10.4	469
10	325	10.7	381
10	327	10.9	453
10	330	11.1	355
10	332	11.4	376
10	335	11.6	307
10	337	12.0	294
10	340	12.2	264
10	343	12.7	253
10	346	12.9	192
10	350	13.5	217
10	353	13.7	140
10	357	14.5	124
10	360	14.7	73
10	366	15.8	77
10	369	15.9	56
10	377	17.5	39
10	379	17.6	25
10	380	19.8	37

Table C-8. Grade 3 Scale-Score Distribution for Mathematics

Grade	Scale score	CSEM	*N*
3	220	12.7	15
3	226	11.1	7
3	232	9.9	6
3	233	10.1	20
3	237	9.1	17
3	238	9.3	38
3	242	8.7	94
3	246	8.2	137
3	249	7.7	59
3	250	7.9	146
3	252	7.4	76
3	253	7.5	209
3	256	7.3	376
3	258	7.0	160
3	259	7.0	319
3	261	6.8	164
3	262	6.8	407
3	264	6.7	650
3	266	6.5	245
3	267	6.5	473
3	268	6.4	251
3	269	6.4	469
3	271	6.3	805
3	273	6.2	296
3	274	6.2	497
3	275	6.2	319
3	276	6.2	533
3	277	6.1	318
3	278	6.1	527
3	279	6.1	363
3	280	6.1	509
3	282	6.0	911
3	284	6.0	991
3	286	6.0	964
3	288	6.0	1,018
3	290	5.9	1,027
3	292	6.0	1,030
3	294	6.0	1,036
3	296	6.0	1,057
3	298	6.0	1,043
3	299	6.8	3
3	300	6.0	1,070
3	301	6.7	6
3	302	6.1	1,101

Grade	Scale score	CSEM	*N*
3	303	6.6	8
3	304	6.1	1,159
3	305	6.6	4
3	306	6.3	1,078
3	307	6.6	16
3	308	6.3	609
3	309	6.3	533
3	310	6.4	642
3	311	6.4	528
3	312	6.5	560
3	313	6.5	606
3	314	6.6	596
3	315	6.6	542
3	316	6.8	74
3	317	6.7	594
3	318	6.8	595
3	319	6.9	581
3	321	6.9	543
3	322	7.0	551
3	323	7.0	500
3	324	6.9	213
3	325	7.2	506
3	326	7.1	481
3	327	7.6	204
3	328	7.2	254
3	329	7.2	427
3	330	7.4	249
3	331	7.5	383
3	332	7.7	94
3	333	7.5	283
3	334	7.5	370
3	335	7.7	261
3	336	7.5	329
3	338	8.0	292
3	339	7.7	259
3	341	8.2	277
3	342	7.8	230
3	343	8.7	11
3	344	8.4	241
3	345	8.0	245
3	346	9.8	3
3	347	8.7	229
3	348	8.3	230
3	350	9.1	228
3	351	8.5	223
3	354	9.5	179
3	355	8.9	177

Grade	Scale score	CSEM	*N*
3	358	9.6	315
3	362	10.2	262
3	367	10.8	227
3	372	11.6	202
3	377	12.2	67
3	378	12.8	94
3	380	13.7	314

Table C-9. Grade 4 Scale-Score Distribution for Mathematics

Grade	Scale score	CSEM	*N*
4	220	10.0	23
4	225	9.2	15
4	231	8.6	24
4	235	8.1	72
4	239	7.7	113
4	243	7.4	234
4	246	7.1	352
4	249	6.9	462
4	252	6.7	612
4	255	6.6	719
4	257	6.5	819
4	260	6.4	897
4	262	6.3	906
4	265	6.2	963
4	267	6.1	983
4	269	6.1	1,036
4	271	6.0	1,021
4	273	6.0	1,100
4	275	5.9	990
4	277	5.9	1,109
4	279	5.9	1,046
4	281	5.9	1,118
4	283	5.9	1,062
4	285	5.9	1,079
4	287	5.9	1,059
4	288	6.8	5
4	289	5.9	978
4	291	5.9	1,051
4	293	6.0	1,092
4	295	6.6	12
4	296	6.0	965
4	297	6.5	18
4	298	6.1	1,004
4	299	6.5	31
4	300	6.1	979
4	301	6.5	60
4	302	6.2	951
4	303	6.4	76
4	304	6.3	932
4	305	6.4	126
4	306	6.3	867
4	307	6.4	161
4	309	6.4	959
4	311	6.6	698

Grade	Scale score	CSEM	*N*
4	312	6.5	281
4	314	6.6	905
4	316	6.7	872
4	318	6.6	436
4	319	7.0	332
4	320	6.7	411
4	322	6.9	747
4	325	7.0	674
4	327	7.0	449
4	328	7.7	80
4	330	7.2	486
4	331	8.1	23
4	332	7.3	467
4	335	7.5	442
4	337	7.7	424
4	339	8.9	2
4	340	8.0	400
4	344	8.2	372
4	347	8.5	362
4	350	8.9	295
4	354	9.3	251
4	359	9.9	243
4	363	10.5	200
4	369	11.4	169
4	375	12.5	154
4	380	14.1	259

Table C-10. Grade 5 Scale-Score Distribution for Mathematics

Grade	Scale score	CSEM	*N*
5	220	12.5	11
5	221	11.1	3
5	223	10.9	1
5	228	10.1	10
5	229	9.9	7
5	233	9.3	40
5	234	9.2	7
5	238	8.7	78
5	239	8.6	32
5	242	8.3	120
5	243	8.2	50
5	245	7.9	253
5	246	7.8	109
5	248	7.6	341
5	250	7.5	187
5	251	7.3	510
5	253	7.2	226
5	254	7.1	595
5	255	7.0	295
5	257	6.9	602
5	258	6.8	327
5	259	6.7	702
5	260	6.6	382
5	262	6.5	681
5	263	6.5	394
5	264	6.4	638
5	265	6.4	430
5	266	6.3	702
5	267	6.2	431
5	268	6.2	640
5	269	6.2	458
5	271	6.1	630
5	272	6.1	477
5	273	6.0	595
5	274	6.0	446
5	275	6.0	605
5	276	5.9	505
5	277	5.9	629
5	278	5.9	1,105
5	279	5.8	471
5	280	5.8	536
5	281	5.8	478
5	282	5.8	567
5	283	5.8	448

Grade	Scale score	CSEM	*N*
5	284	5.8	525
5	285	5.8	444
5	286	5.8	502
5	287	5.8	482
5	288	5.8	541
5	289	5.8	526
5	290	5.8	494
5	291	5.9	443
5	292	5.9	552
5	293	5.9	480
5	294	6.0	495
5	295	5.9	482
5	296	6.0	563
5	297	6.0	459
5	298	6.0	507
5	299	6.0	416
5	300	6.0	455
5	301	6.1	451
5	302	6.1	425
5	303	6.0	403
5	304	6.1	388
5	305	6.0	398
5	306	6.1	391
5	307	6.0	368
5	308	6.1	383
5	309	6.0	354
5	310	6.1	327
5	311	6.0	329
5	312	6.2	329
5	313	6.0	304
5	314	6.2	319
5	315	6.1	293
5	316	6.2	562
5	317	6.7	2
5	318	6.2	492
5	320	6.3	507
5	322	6.3	499
5	324	6.5	246
5	325	6.3	218
5	326	6.6	203
5	327	6.4	204
5	328	6.7	212
5	329	6.5	206
5	331	6.7	401
5	333	6.8	388
5	336	7.0	396
5	338	7.2	318

Grade	Scale score	CSEM	*N*
5	341	7.4	319
5	344	7.8	277
5	347	8.1	277
5	350	8.1	119
5	351	8.7	133
5	353	8.5	130
5	355	9.2	119
5	357	9.1	94
5	359	9.8	86
5	361	9.8	84
5	364	10.6	88
5	367	10.7	72
5	369	11.7	75
5	373	11.9	49
5	376	13.1	71
5	380	14.6	181

Table C-11. Grade 6 Scale-Score Distribution for Mathematics

Grade	Scale score	CSEM	*N*
6	220	15.6	15
6	221	13.1	5
6	228	11.5	11
6	233	10.4	15
6	234	10.4	7
6	238	9.5	33
6	239	9.5	20
6	243	8.9	126
6	246	8.3	157
6	247	8.4	96
6	250	7.9	427
6	253	7.6	390
6	254	7.6	218
6	256	7.2	808
6	258	7.0	643
6	259	7.0	341
6	261	6.8	722
6	262	6.8	406
6	263	6.6	763
6	264	6.6	451
6	266	6.4	1,265
6	268	6.3	800
6	269	6.3	499
6	270	6.2	774
6	271	6.2	488
6	272	6.1	690
6	273	6.1	516
6	274	6.0	640
6	275	6.0	535
6	276	5.9	604
6	277	6.0	513
6	278	5.9	604
6	279	5.9	465
6	280	5.9	585
6	281	5.9	466
6	282	5.8	534
6	283	5.9	1,009
6	284	5.9	483
6	285	5.9	548
6	286	5.9	475
6	287	5.9	546
6	288	5.9	509
6	289	5.9	526
6	290	6.0	440

Grade	Scale score	CSEM	*N*
6	291	6.0	524
6	292	6.0	503
6	293	6.0	518
6	294	6.1	473
6	295	6.1	509
6	296	6.1	435
6	297	6.2	514
6	298	6.2	438
6	299	6.3	495
6	300	6.2	488
6	301	6.3	492
6	302	6.4	455
6	303	6.3	430
6	304	6.4	467
6	305	6.4	423
6	306	6.4	424
6	307	6.3	278
6	308	6.6	388
6	309	6.4	355
6	310	6.9	50
6	311	6.5	614
6	312	7.1	41
6	313	6.6	293
6	314	6.5	305
6	315	6.7	323
6	316	6.5	299
6	317	6.8	308
6	318	6.6	275
6	319	6.9	287
6	320	6.7	272
6	321	7.8	4
6	322	6.9	502
6	324	7.2	259
6	325	6.9	197
6	327	7.1	448
6	329	7.5	220
6	330	7.1	190
6	332	7.5	380
6	335	7.7	382
6	338	8.0	369
6	341	8.3	315
6	344	8.6	308
6	347	8.5	148
6	348	9.4	116
6	351	8.9	114
6	352	10.0	118
6	355	9.5	107

Grade	Scale score	CSEM	*N*
6	357	10.7	113
6	360	10.3	102
6	362	11.5	85
6	365	11.3	100
6	368	12.7	81
6	372	12.7	69
6	375	14.2	69
6	380	15.6	284

Table C-12. Grade 7 Scale-Score Distribution for Mathematics

Grade	Scale score	CSEM	*N*
7	220	12.0	39
7	221	11.2	21
7	222	11.2	8
7	227	10.5	35
7	228	10.5	15
7	232	9.9	76
7	233	9.9	21
7	237	9.5	169
7	241	9.1	291
7	245	8.8	449
7	248	8.5	444
7	249	8.4	243
7	252	8.2	844
7	255	8.0	1,083
7	258	7.9	1,191
7	261	7.7	1,302
7	264	7.5	1,386
7	267	7.4	1,411
7	269	7.3	1,493
7	272	7.2	1,471
7	274	7.1	1,442
7	277	7.0	1,448
7	279	7.0	1,321
7	280	7.9	2
7	281	6.9	1,307
7	282	7.5	1
7	283	7.8	1
7	284	6.9	1,245
7	286	6.9	1,146
7	288	6.9	1,169
7	290	6.8	541
7	291	7.0	596
7	292	7.2	5
7	293	6.9	1,041
7	295	6.9	985
7	297	6.9	528
7	298	7.1	479
7	299	6.9	461
7	300	7.1	476
7	302	7.1	521
7	303	7.2	365
7	304	7.1	464
7	305	7.3	455
7	306	7.2	125

Grade	Scale score	CSEM	*N*
7	307	7.3	408
7	308	7.4	375
7	309	7.3	412
7	310	7.5	181
7	311	7.3	173
7	312	7.4	430
7	313	7.5	323
7	314	7.5	237
7	315	7.6	110
7	316	7.5	572
7	318	7.6	321
7	319	7.6	245
7	321	7.7	533
7	322	8.1	33
7	323	7.7	227
7	324	7.8	249
7	325	8.4	6
7	326	7.9	246
7	327	7.9	221
7	328	8.0	230
7	329	8.1	221
7	331	8.2	189
7	332	8.2	206
7	334	8.4	167
7	335	8.4	184
7	337	8.6	158
7	338	8.6	183
7	339	9.9	1
7	340	8.8	162
7	341	8.8	143
7	344	9.0	287
7	347	9.3	129
7	348	9.3	117
7	351	9.6	214
7	355	9.9	190
7	359	10.3	78
7	360	10.4	94
7	363	10.8	61
7	364	10.8	71
7	368	11.4	52
7	369	11.4	63
7	374	12.1	51
7	375	12.1	38
7	380	13.1	204

Table C-13. Grade 8 Scale-Score Distribution for Mathematics

Grade	Scale score	CSEM	*N*
8	220	11.7	85
8	221	10.9	43
8	226	10.2	62
8	227	10.3	28
8	231	9.7	164
8	236	9.3	257
8	240	8.9	390
8	244	8.6	594
8	247	8.3	805
8	250	8.1	1,034
8	253	7.9	1,188
8	256	7.7	1,343
8	259	7.6	1,414
8	262	7.4	1,435
8	265	7.3	1,532
8	267	7.2	1,498
8	268	8.3	1
8	270	7.1	1,403
8	271	8.1	11
8	272	7.1	1,417
8	274	7.9	19
8	275	7.0	1,364
8	277	7.0	1,399
8	280	7.0	1,296
8	282	7.0	1,186
8	283	7.6	66
8	284	7.0	488
8	285	7.2	790
8	287	7.0	958
8	288	7.4	321
8	289	7.0	795
8	290	7.4	410
8	292	7.2	818
8	293	7.3	256
8	294	7.1	470
8	295	7.3	555
8	297	7.2	646
8	298	7.2	338
8	299	7.3	200
8	300	7.2	641
8	302	7.2	763
8	304	7.4	34
8	305	7.2	695
8	307	7.3	650

Grade	Scale score	CSEM	*N*
8	308	7.6	10
8	309	7.2	286
8	310	7.3	280
8	312	7.4	558
8	313	7.7	2
8	314	7.3	250
8	315	7.4	263
8	317	7.4	515
8	319	7.5	217
8	320	7.6	226
8	322	7.6	235
8	323	7.7	235
8	325	7.7	384
8	327	7.9	167
8	328	7.9	164
8	330	8.0	206
8	331	8.1	176
8	333	8.2	168
8	334	8.2	173
8	336	8.5	146
8	337	8.4	167
8	340	8.7	290
8	343	9.1	123
8	344	8.8	142
8	347	9.2	263
8	351	9.6	209
8	355	10.0	193
8	360	10.4	184
8	364	10.4	67
8	365	11.6	78
8	370	11.0	61
8	371	12.4	55
8	375	11.9	64
8	378	13.5	50
8	380	14.1	291

Table C-14. Grade 10 Scale-Score Distribution for Mathematics

Grade	Scale score	CSEM	*N*
10	220	12.0	172
10	224	11.1	33
10	225	11.0	16
10	230	10.4	80
10	235	9.8	120
10	239	9.3	234
10	243	8.9	348
10	247	8.5	546
10	250	8.2	776
10	253	8.0	578
10	254	7.9	414
10	256	7.7	716
10	257	7.7	501
10	259	7.5	1,340
10	262	7.3	1,534
10	265	7.2	1,551
10	267	7.1	1,540
10	270	6.9	1,618
10	272	6.9	1,607
10	273	7.7	5
10	274	6.8	1,447
10	276	7.5	2
10	277	6.7	1,387
10	278	7.5	5
10	279	6.7	1,375
10	281	6.6	1,274
10	283	6.6	1,207
10	286	6.6	1,163
10	288	6.7	1,030
10	290	6.6	953
10	291	7.1	43
10	292	6.8	835
10	293	7.1	70
10	295	6.8	889
10	297	6.9	682
10	298	7.1	146
10	299	6.9	600
10	300	7.1	201
10	302	7.1	571
10	303	7.2	123
10	304	7.1	343
10	305	7.2	322
10	307	7.3	606
10	309	7.4	87

Grade	Scale score	CSEM	*N*
10	310	7.4	460
10	312	7.5	529
10	315	7.6	461
10	317	7.6	431
10	318	8.0	31
10	320	7.8	392
10	321	8.2	10
10	322	7.9	217
10	323	7.9	230
10	324	8.5	3
10	325	8.1	385
10	328	8.2	358
10	331	8.3	345
10	334	8.6	320
10	337	8.8	152
10	338	8.8	154
10	341	9.0	292
10	344	9.3	236
10	348	9.6	257
10	352	9.9	233
10	356	10.3	217
10	360	10.6	82
10	361	10.7	89
10	365	11.1	74
10	366	11.2	95
10	371	11.7	119
10	376	12.4	64
10	377	12.6	48
10	380	13.5	279

Table C-15. Grade 5 Scale-Score Distribution for Science

Grade	Scale score	CSEM	*N*
5	220	15.7	121
5	222	15.2	12
5	223	15.2	18
5	227	14.0	120
5	230	13.8	22
5	231	13.8	25
5	234	12.9	164
5	237	12.8	91
5	240	12.2	249
5	243	12.1	133
5	245	11.6	65
5	246	11.7	272
5	248	11.5	85
5	249	11.5	110
5	250	11.2	98
5	251	11.3	336
5	253	11.0	134
5	254	11.1	135
5	255	11.0	572
5	258	10.8	333
5	260	10.7	705
5	263	10.6	427
5	264	10.5	723
5	267	10.4	513
5	268	10.4	852
5	271	10.3	553
5	272	10.3	315
5	273	10.4	632
5	275	10.2	727
5	276	10.2	378
5	277	10.4	612
5	279	10.2	787
5	281	10.4	1,039
5	283	10.3	799
5	285	10.4	1,113
5	287	10.4	931
5	289	10.6	1,189
5	291	10.5	997
5	293	10.8	1,225
5	295	10.5	483
5	296	10.8	520
5	297	10.8	490
5	298	11.1	708
5	299	10.8	479

Grade	Scale score	CSEM	*N*
5	300	11.0	534
5	302	11.3	1,285
5	304	11.3	1,068
5	306	11.5	527
5	307	11.9	639
5	309	11.7	1,026
5	311	12.0	533
5	312	12.4	615
5	314	12.1	1,034
5	317	12.6	519
5	318	13.0	561
5	319	12.6	484
5	320	13.0	441
5	323	13.4	487
5	324	13.8	506
5	325	13.4	500
5	326	13.8	457
5	329	14.4	459
5	331	14.7	879
5	333	14.8	367
5	337	15.7	375
5	338	16.2	398
5	339	15.7	325
5	340	16.1	332
5	346	17.4	299
5	347	17.6	548
5	349	17.9	235
5	356	19.7	225
5	358	20.0	404
5	360	20.3	195
5	369	23.1	148
5	371	23.4	283
5	373	23.7	127
5	380	28.7	512

Table C-16. Grade 8 Scale-Score Distribution for Science

Grade	Scale score	CSEM	*N*
8	220	15.7	233
8	224	14.8	68
8	225	14.4	22
8	226	14.7	102
8	230	14.1	149
8	232	13.8	200
8	236	13.4	238
8	238	13.2	310
8	242	12.9	362
8	243	12.7	406
8	247	12.5	424
8	248	12.2	511
8	251	12.2	306
8	252	12.0	793
8	253	11.8	165
8	256	11.8	683
8	257	11.6	682
8	260	11.7	424
8	261	11.4	901
8	262	11.3	212
8	264	11.5	509
8	265	11.3	1,030
8	266	11.1	235
8	268	11.4	542
8	269	11.2	1,003
8	270	11.0	259
8	272	11.3	552
8	273	11.1	1,007
8	274	10.9	275
8	276	11.2	604
8	277	11.1	1,079
8	278	10.9	307
8	280	11.1	1,112
8	281	11.2	628
8	282	10.8	301
8	284	11.1	1,055
8	285	11.2	616
8	286	10.9	314
8	288	11.1	999
8	289	11.3	595
8	290	10.9	310
8	292	11.3	949
8	293	11.4	629
8	294	11.0	289

Grade	Scale score	CSEM	*N*
8	296	11.4	882
8	298	11.4	921
8	300	11.6	913
8	302	11.6	866
8	304	11.8	856
8	306	11.8	852
8	308	11.9	290
8	309	12.2	537
8	311	12.1	749
8	313	12.5	739
8	315	12.2	253
8	316	12.6	452
8	318	12.9	663
8	320	12.6	229
8	321	13.0	405
8	323	13.2	201
8	324	13.6	397
8	326	13.2	153
8	327	13.6	403
8	329	14.1	512
8	332	13.8	129
8	333	14.2	304
8	335	14.6	174
8	336	15.1	296
8	338	14.5	142
8	339	15.0	241
8	342	15.6	103
8	343	16.1	219
8	345	15.5	119
8	347	15.9	184
8	349	16.9	102
8	351	17.5	179
8	354	16.7	81
8	355	17.2	150
8	359	18.6	63
8	360	19.2	115
8	364	18.3	47
8	365	18.8	95
8	370	21.1	46
8	372	21.7	69
8	376	20.6	33
8	378	21.1	50
8	380	25.0	170

Table C-17. Grade 11 Scale-Score Distribution for Science

Grade	Scale score	CSEM	*N*
11	220	16.8	82
11	223	14.8	8
11	224	14.7	13
11	225	14.5	25
11	226	14.4	18
11	230	13.2	16
11	231	13.2	32
11	232	13.0	43
11	233	12.9	38
11	237	12.0	102
11	238	11.9	81
11	239	11.8	79
11	242	11.1	58
11	243	11.0	262
11	244	11.1	134
11	247	10.4	88
11	248	10.4	392
11	249	10.5	193
11	251	9.9	119
11	252	10.0	516
11	253	10.0	298
11	255	9.5	161
11	256	9.6	644
11	257	9.7	326
11	259	9.2	168
11	260	9.3	691
11	261	9.4	368
11	263	9.1	951
11	264	9.2	415
11	266	8.9	252
11	267	9.0	784
11	268	9.1	428
11	269	8.8	233
11	270	8.9	796
11	271	9.0	450
11	273	8.8	554
11	274	9.0	955
11	276	8.9	547
11	277	8.9	910
11	279	8.8	250
11	280	8.9	854
11	281	9.0	456
11	282	8.9	275
11	283	9.0	273

Grade	Scale score	CSEM	*N*
11	284	9.0	952
11	286	9.1	560
11	287	9.2	1,002
11	289	9.3	529
11	290	9.3	533
11	291	9.4	496
11	293	9.5	523
11	294	9.5	1,023
11	296	9.7	568
11	297	9.8	533
11	298	9.8	527
11	300	10.0	545
11	301	10.1	1,023
11	304	10.4	531
11	305	10.4	999
11	308	10.8	451
11	309	10.8	939
11	312	11.3	218
11	313	11.3	676
11	314	11.3	429
11	317	11.8	404
11	318	11.9	408
11	319	11.9	427
11	322	12.6	231
11	323	12.5	597
11	324	12.5	395
11	328	13.4	382
11	329	13.3	347
11	330	13.3	362
11	334	14.4	154
11	335	14.3	463
11	336	14.2	292
11	341	15.7	118
11	342	15.5	379
11	343	15.3	242
11	349	17.3	89
11	350	17.0	287
11	351	16.7	202
11	359	19.2	234
11	360	18.7	57
11	361	18.4	115
11	370	22.2	41
11	371	21.8	90
11	372	20.9	138
11	380	25.1	312

XI. Appendix D: Subgroup Reliability and Performance

For all tables in Appendix E: NHPI = Native Hawaiian and Pacific Islander; SWD = students with disabilities; EL = English learners.

Table D-1. Grade 3 Subgroup Reliability and Performance for ELA

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	35,902					
Black		2,836	7.9%	0.93	279.0	25.9
American Indian		1,033	2.9%	0.93	283.6	24.9
Asian		1,089	3.0%	0.92	303.8	30.7
NHPI		91	0.3%	0.93	290.3	28.6
White		30,853	85.9%	0.92	297.3	28.6
Hispanic	38,340					
Yes		7,790	20.3%	0.93	283.4	25.1
No		30,550	79.7%	0.92	298.5	28.9
SWD	38,340					
Yes		4,879	12.7%	0.92	274.9	26.0
No		33,461	87.3%	0.92	298.4	28.0
ELL	38,340					
Yes		5,118	13.3%	0.93	279.4	23.7
No		33,222	86.7%	0.92	297.9	28.8

Table D-2. Grade 4 Subgroup Reliability and Performance for ELA

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	35,929					
Black		2,789	7.8%	0.91	283.6	25.6
American Indian		998	2.8%	0.91	289.1	24.1
Asian		1,117	3.1%	0.89	310.6	30.3
NHPI		89	0.2%	0.91	289.7	27.2
White		30,936	86.1%	0.90	302.3	27.6
Hispanic	38,424					
Yes		7,629	19.9%	0.91	289.6	25.0
No		30,795	80.1%	0.90	303.2	28.0
SWD	38,424					
Yes		4,912	12.8%	0.91	278.9	25.4
No		33,512	87.2%	0.90	303.7	26.9
ELL	38,424					
Yes		5,022	13.1%	0.92	285.3	23.4
No		33,402	86.9%	0.90	302.8	27.9

Table D-3. Grade 5 Subgroup Reliability and Performance for ELA

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	35,088					
Black		2,711	7.7%	0.92	279.1	27.1
American Indian		1,122	3.2%	0.92	284.9	25.8
Asian		1,070	3.0%	0.90	307.1	31.3
NHPI		85	0.2%	0.92	287.3	27.7
White		30,100	85.8%	0.91	298.9	29.6
Hispanic	37,526					
Yes		7,491	20.0%	0.92	285.2	26.7
No		30,035	80.0%	0.91	299.9	29.9
SWD	37,526					
Yes		4,722	12.6%	0.92	271.9	25.4
No		32,804	87.4%	0.91	300.6	28.7
ELL	37,526					
Yes		4,820	12.8%	0.92	280.5	25.1
No		32,706	87.3%	0.91	299.4	29.8

Table D-4. Grade 6 Subgroup Reliability and Performance for ELA

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,482					
Black		2,572	7.5%	0.92	273.1	26.4
American Indian		1,326	3.8%	0.92	278.0	26.0
Asian		1,038	3.0%	0.90	303.0	30.3
NHPI		88	0.3%	0.92	282.9	29.2
White		29,458	85.4%	0.91	293.2	28.4
Hispanic	36,858					
Yes		7,220	19.6%	0.92	278.5	26.4
No		29,638	80.4%	0.91	294.2	28.6
SWD	36,858					
Yes		4,440	12.0%	0.92	264.5	25.2
No		32,418	88.0%	0.91	294.8	27.4
ELL	36,858					
Yes		4,645	12.6%	0.92	274.2	25.7
No		32,213	87.4%	0.91	293.6	28.5

Table D-5. Grade 7 Subgroup Reliability and Performance for ELA

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,566					
Black		2,523	7.3%	0.92	272.3	26.4
American Indian		1,427	4.1%	0.92	276.1	25.8
Asian		1,116	3.2%	0.89	299.1	33.2
NHPI		106	0.3%	0.92	274.9	27.8
White		29,394	85.0%	0.90	291.8	30.5
Hispanic	36,863					
Yes		7,167	19.4%	0.92	277.4	26.6
No		29,696	80.6%	0.90	292.7	30.9
SWD	36,863					
Yes		4,240	11.5%	0.93	261.2	22.6
No		32,623	88.5%	0.90	293.4	29.6
ELL	36,863					
Yes		4,483	12.2%	0.92	272.0	24.5
No		32,380	87.8%	0.90	292.2	30.6

Table D-6. Grade 8 Subgroup Reliability and Performance for ELA

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score *M*	Scaled score *SD*
Race	34,572					
Black		2,568	7.4%	0.91	267.0	24.3
American Indian		1,483	4.3%	0.91	271.5	24.0
Asian		1,101	3.2%	0.89	294.8	30.7
NHPI		87	0.3%	0.90	282.9	27.8
White		29,333	84.8%	0.90	285.9	28.1
Hispanic	36,695					
Yes		6,906	18.8%	0.91	272.8	25.4
No		29,789	81.2%	0.90	286.5	28.4
SWD	36,695					
Yes		4,066	11.1%	0.91	256.1	21.3
No		32,629	88.9%	0.90	287.4	27.2
ELL	36,695					
Yes		4,287	11.7%	0.91	267.2	23.1
No		32,408	88.3%	0.90	286.2	28.3

Table D-7. Grade 10 Subgroup Reliability and Performance for ELA

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score *M*	Scaled score *SD*
Race	33,663					
Black		2,314	6.9%	0.93	266.8	26.1
American Indian		1,365	4.1%	0.93	269.9	25.1
Asian		1,136	3.4%	0.91	291.6	32.4
NHPI		77	0.2%	0.93	273.0	26.0
White		28,771	85.5%	0.91	286.9	29.5
Hispanic	35,673					
Yes		6,371	17.9%	0.93	272.2	26.7
No		29,302	82.1%	0.91	287.5	29.7
SWD	35,673					
Yes		3,613	10.1%	0.93	256.3	22.1
No		32,060	89.9%	0.91	288.0	28.8
ELL	35,673					
Yes		3,601	10.1%	0.93	264.5	24.4
No		32,072	89.9%	0.91	287.1	29.5

Table D-8. Grade 3 Subgroup Reliability and Performance for Mathematics

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	36,003					
Black		2,847	7.9%	0.94	287.0	24.3
American Indian		1,040	2.9%	0.94	294.0	25.1
Asian		1,109	3.1%	0.92	315.9	31.7
NHPI		92	0.3%	0.94	298.1	25.8
White		30,915	85.9%	0.94	304.9	27.1
Hispanic	38,438					
Yes		7,867	20.5%	0.94	292.4	23.6
No		30,571	79.5%	0.94	306.0	27.7
SWD	38,438					
Yes		4,870	12.7%	0.94	284.7	25.9
No		33,568	87.3%	0.94	305.9	26.7
ELL	38,438					
Yes		5,261	13.7%	0.94	290.3	24.0
No		33,177	86.3%	0.94	305.2	27.4

Table D-9. Grade 4 Subgroup Reliability and Performance for Mathematics

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	36,022					
Black		2,810	7.8%	0.95	275.1	22.8
American Indian		1,004	2.8%	0.95	283.0	23.5
Asian		1,139	3.2%	0.93	308.7	33.2
NHPI		89	0.2%	0.95	281.8	26.5
White		30,980	86.0%	0.94	295.8	27.8
Hispanic	38,514					
Yes		7,693	20.0%	0.95	282.9	23.9
No		30,821	80.0%	0.94	296.7	28.5
SWD	38,514					
Yes		4,906	12.7%	0.95	274.4	24.4
No		33,608	87.3%	0.94	296.8	27.6
ELL	38,514					
Yes		5,146	13.4%	0.95	280.5	23.7
No		33,368	86.6%	0.94	296.0	28.3

Table D-10. Grade 5 Subgroup Reliability and Performance for Mathematics

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	35,169					
Black		2,733	7.8%	0.94	274.1	21.3
American Indian		1,129	3.2%	0.95	280.9	22.2
Asian		1,091	3.1%	0.93	306.5	33.7
NHPI		85	0.2%	0.95	285.8	22.4
White		30,131	85.7%	0.94	292.7	27.4
Hispanic	37,608					
Yes		7,555	20.1%	0.95	280.3	22.9
No		30,053	79.9%	0.94	293.8	28.0
SWD	37,608					
Yes		4,711	12.5%	0.94	270.5	22.3
No		32,897	87.5%	0.94	294.0	27.0
ELL	37,608					
Yes		4,939	13.1%	0.95	277.9	22.7
No		32,669	86.9%	0.94	293.1	27.7

Table D-11. Grade 6 Subgroup Reliability and Performance for Mathematics

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,551					
Black		2,584	7.5%	0.94	275.3	19.7
American Indian		1,333	3.9%	0.94	280.6	21.5
Asian		1,059	3.1%	0.91	308.8	33.4
NHPI		89	0.3%	0.94	284.7	25.1
White		29,486	85.3%	0.94	292.9	26.6
Hispanic	36,923					
Yes		7,272	19.7%	0.94	280.2	21.7
No		29,651	80.3%	0.93	294.0	27.3
SWD	36,923					
Yes		4,432	12.0%	0.94	270.6	19.5
No		32,491	88.0%	0.94	294.1	26.4
ELL	36,923					
Yes		4,746	12.9%	0.94	278.4	21.6
No		32,177	87.1%	0.94	293.2	27.0

Table D-12. Grade 7 Subgroup Reliability and Performance for Mathematics

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,614					
Black		2,536	7.3%	0.92	271.2	20.6
American Indian		1,434	4.1%	0.93	276.6	22.0
Asian		1,126	3.3%	0.91	305.9	35.4
NHPI		105	0.3%	0.93	276.9	22.3
White		29,413	85.0%	0.92	290.1	27.3
Hispanic	36,910					
Yes		7,221	19.6%	0.93	277.3	22.6
No		29,689	80.4%	0.92	292.1	28.1
SWD	36,910					
Yes		4,227	11.5%	0.92	264.1	18.8
No		32,683	88.5%	0.93	291.5	27.1
ELL	36,910					
Yes		4,572	12.4%	0.92	274.5	22.2
No		32,338	87.6%	0.92	290.3	27.8

Table D-13. Grade 8 Subgroup Reliability and Performance for Mathematics

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,635					
Black		2,575	7.4%	0.93	267.7	21.8
American Indian		1,487	4.3%	0.93	273.1	22.2
Asian		1,113	3.2%	0.92	306.3	37.2
NHPI		90	0.3%	0.93	282.0	28.5
White		29,370	84.8%	0.93	286.4	28.7
Hispanic	36,758					
Yes		6,954	18.9%	0.93	273.9	24.4
No		29,804	81.1%	0.93	287.3	29.5
SWD	36,758					
Yes		4,059	11.0%	0.93	259.7	18.8
No		32,699	89.0%	0.93	287.8	28.6
ELL	36,758					
Yes		4,382	11.9%	0.93	271.2	23.4
No		32,376	88.1%	0.93	286.6	29.3

Table D-14. Grade 10 Subgroup Reliability and Performance for Mathematics

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score *M*	Scaled score *SD*
Race	33,651					
Black		2,305	6.8%	0.93	269.4	21.6
American Indian		1,369	4.1%	0.93	272.1	18.8
Asian		1,146	3.4%	0.92	305.3	36.2
NHPI		77	0.2%	0.93	275.0	22.1
White		28,754	85.4%	0.93	287.0	28.3
Hispanic	35,653					
Yes		6,386	17.9%	0.93	273.8	22.0
No		29,267	82.1%	0.93	288.1	29.1
SWD	35,653					
Yes		3,589	10.1%	0.93	263.1	17.7
No		32,064	89.9%	0.93	288.1	28.4
ELL	35,653					
Yes		3,643	10.2%	0.93	270.6	21.1
No		32,010	89.8%	0.93	287.3	28.7

Table D-15. Grade 5 Subgroup Reliability and Performance for Science

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score *M*	Scaled score *SD*
Race	35,173					
Black		2,727	7.8%	0.86	280.2	25.9
American Indian		1,130	3.2%	0.85	286.9	27.6
Asian		1,092	3.1%	0.78	307.4	33.5
NHPI		85	0.2%	0.85	289.7	27.6
White		30,138	85.7%	0.82	300.5	29.9
Hispanic	37,609					
Yes		7,549	20.1%	0.85	286.8	26.9
No		30,060	79.9%	0.81	301.4	30.3
SWD	37,609					
Yes		4,716	12.5%	0.85	278.4	28.1
No		32,893	87.5%	0.82	301.4	29.4
ELL	37,609					
Yes		4,928	13.1%	0.85	282.4	26.3
No		32,681	86.9%	0.82	300.9	30.0

Table D-16. Grade 8 Subgroup Reliability and Performance for HGSS

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,638					
Black		2,570	7.4%	0.84	268.6	23.4
American Indian		1,491	4.3%	0.84	275.4	24.7
Asian		1,110	3.2%	0.81	295.0	31.7
NHPI		90	0.3%	0.84	278.0	25.1
White		29,377	84.8%	0.83	290.9	29.3
Hispanic	36,766					
Yes		6,963	18.9%	0.84	275.8	25.8
No		29,803	81.1%	0.82	291.4	29.6
SWD	36,766					
Yes		4,053	11.0%	0.83	266.0	24.7
No		32,713	89.0%	0.83	291.2	28.9
ELL	36,766					
Yes		4,384	11.9%	0.84	270.5	23.7
No		32,382	88.1%	0.82	290.8	19.5

Table D-17. Grade 11 Subgroup Reliability and Performance for HGSS

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	32,316					
Black		2,147	6.6%	0.88	272.9	23.1
American Indian		1,366	4.2%	0.88	278.8	24.3
Asian		1,088	3.4%	0.84	295.1	31.8
NHPI		80	0.2%	0.86	288.3	29.9
White		27,635	85.5%	0.85	294.0	28.9
Hispanic	34,158					
Yes		5,878	17.2%	0.88	279.4	25.4
No		28,280	82.8%	0.85	294.3	29.1
SWD	34,158					
Yes		3,278	9.6%	0.88	269.3	21.9
No		30,880	90.4%	0.85	294.1	28.7
ELL	34,158					
Yes		3,058	9.0%	0.88	271.3	21.6
No		31,100	91.0%	0.85	293.7	28.9

XII. Appendix E: Path Reliability

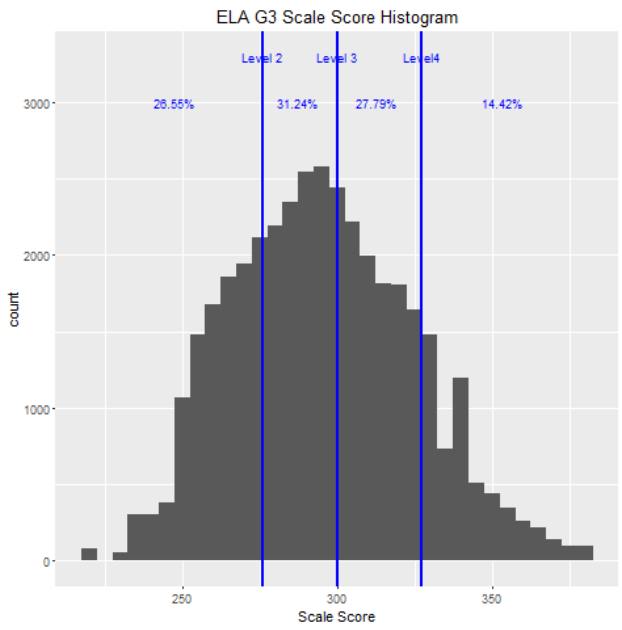
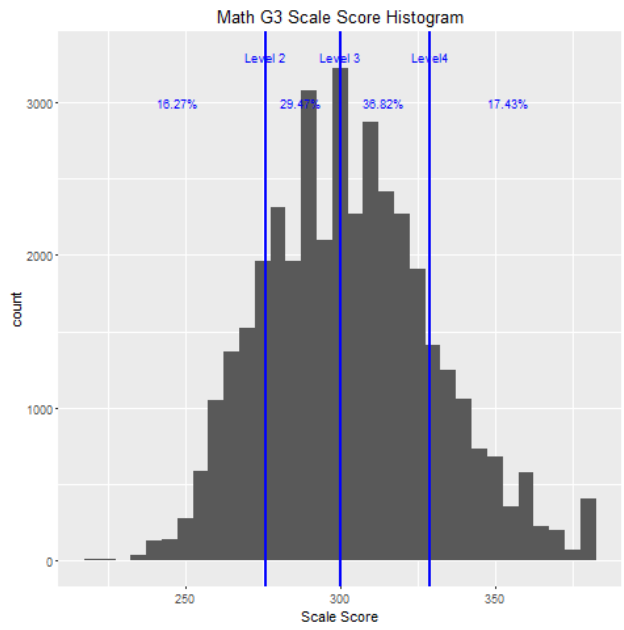
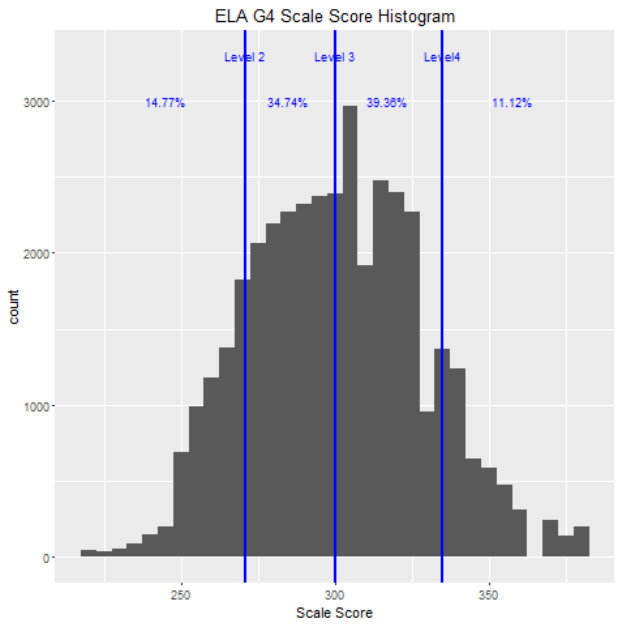
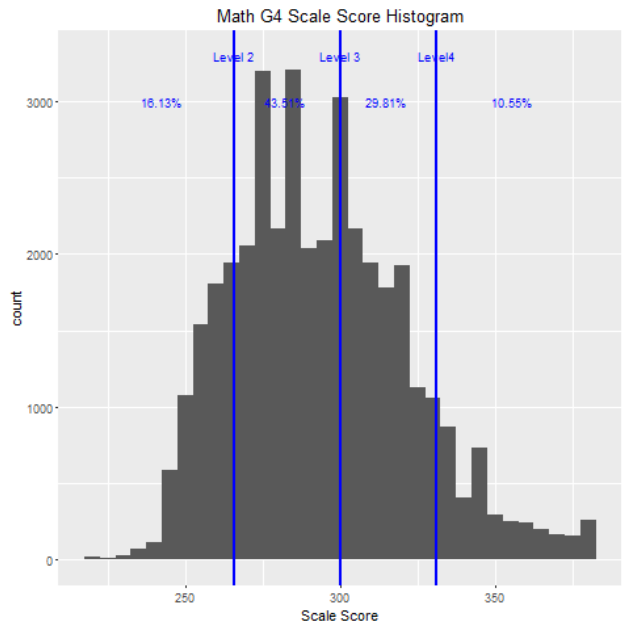
Table E-1. Path Reliability for ELA

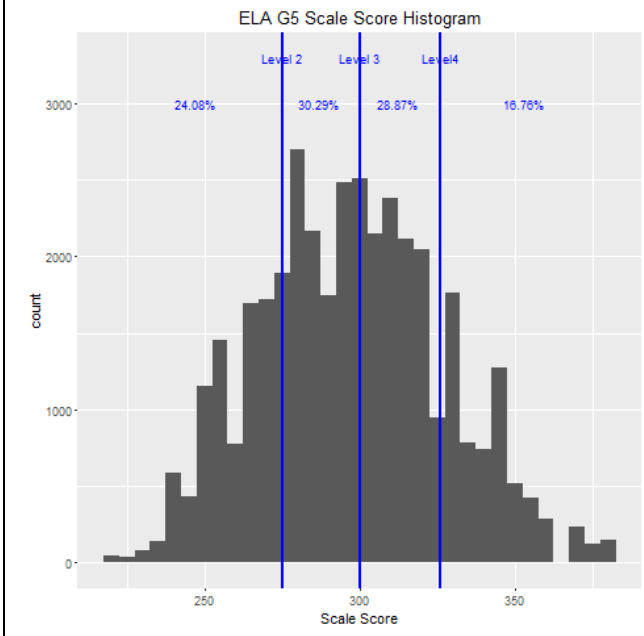
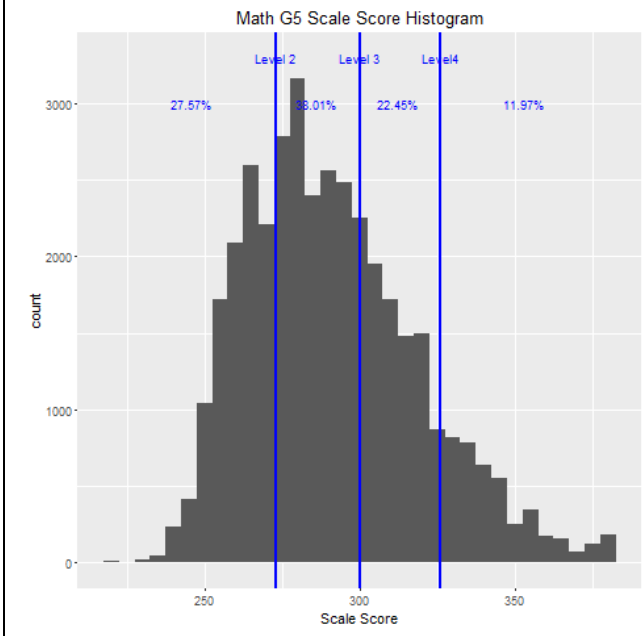
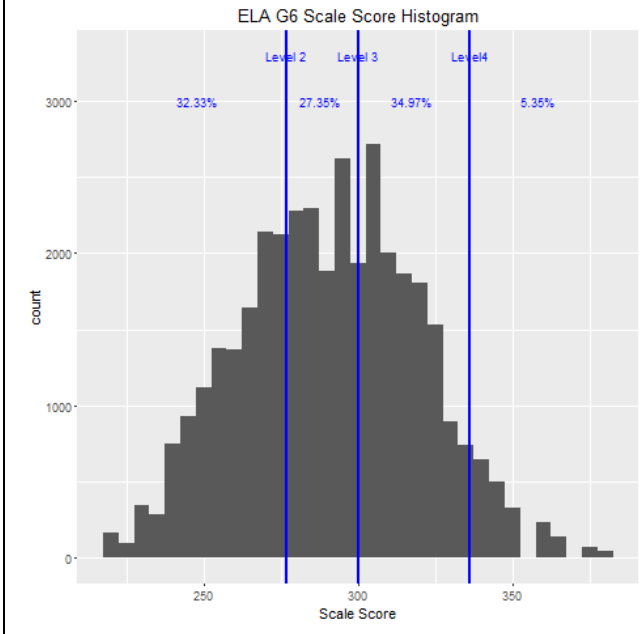
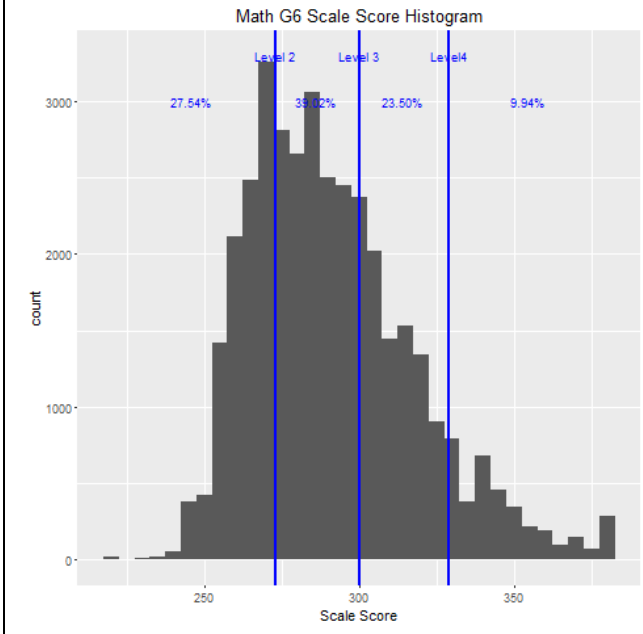
Grade	Path	Stage 1	Stage 2	*N*	%	Reliability
3	1	Medium	Easy	20,360	53.1%	0.93
3	2	Medium	Hard	17,954	46.9%	0.92
4	1	Medium	Easy	7,475	19.5%	0.92
4	2	Medium	Hard	30,902	80.5%	0.90
5	1	Medium	Easy	18,463	49.3%	0.93
5	2	Medium	Hard	19,026	50.8%	0.90
6	1	Medium	Easy	7,118	19.3%	0.92
6	2	Medium	Hard	29,704	80.7%	0.91
7	1	Medium	Easy	19,703	53.5%	0.93
7	2	Medium	Hard	17,103	46.5%	0.87
8	1	Medium	Easy	18,448	50.4%	0.92
8	2	Medium	Hard	18,181	49.6%	0.88
10	1	Medium	Easy	16,610	46.9%	0.94
10	2	Medium	Hard	18,844	53.2%	0.90

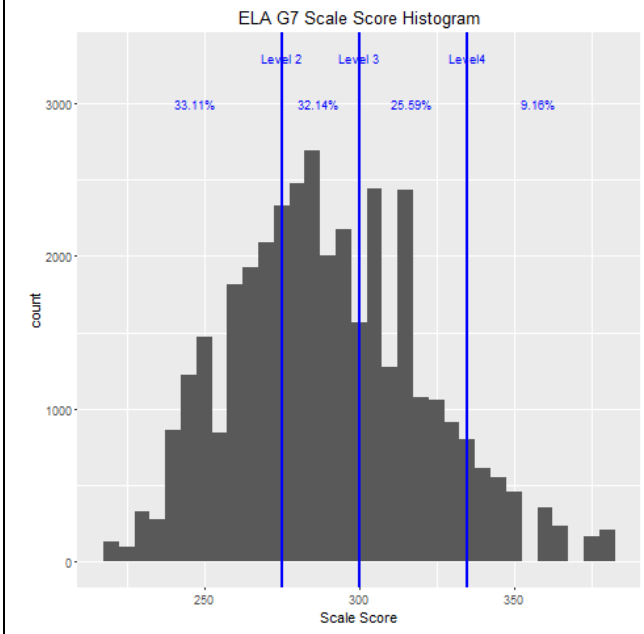
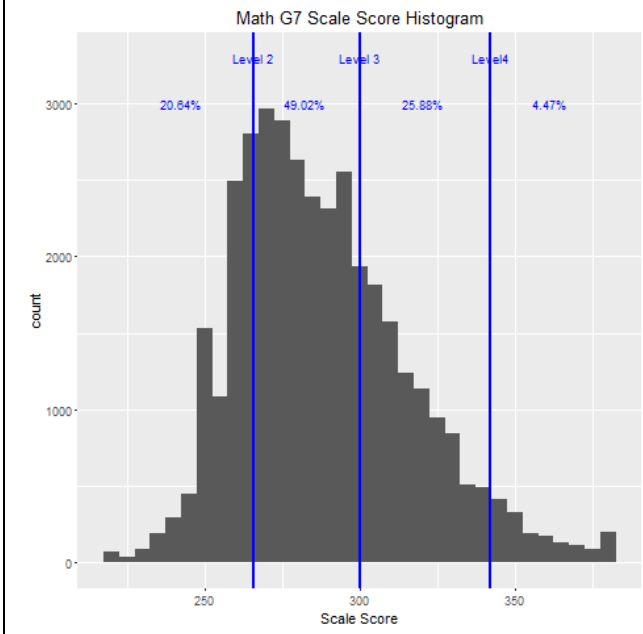
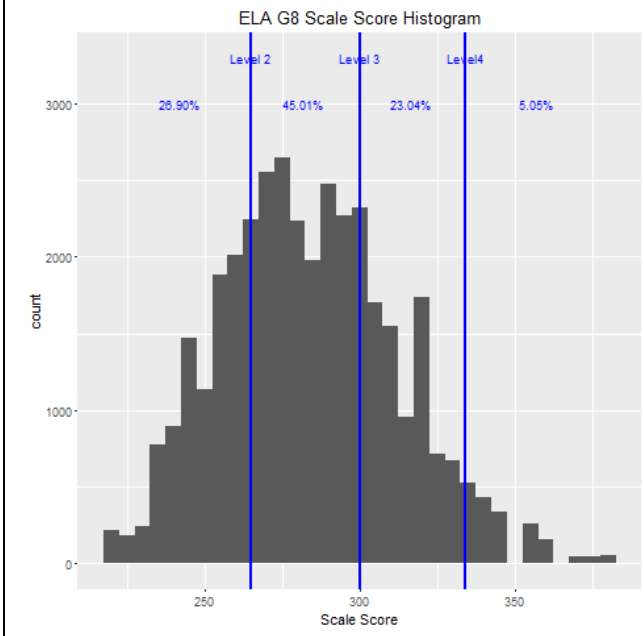
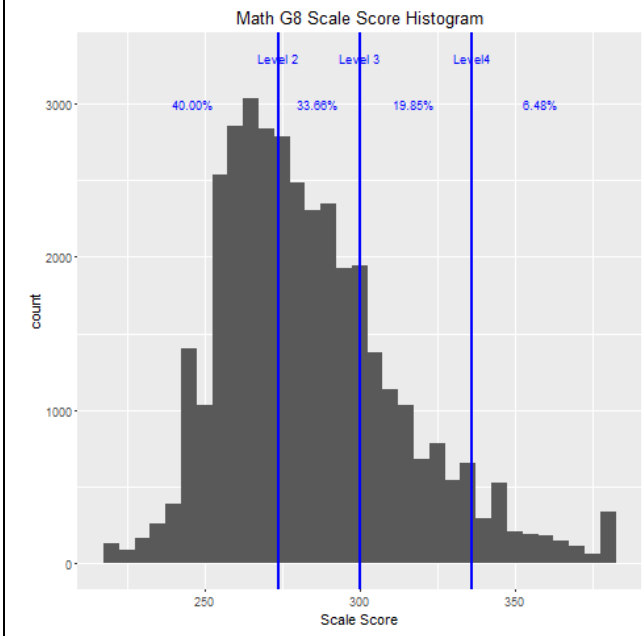
Table E-2. Path Reliability for Mathematics

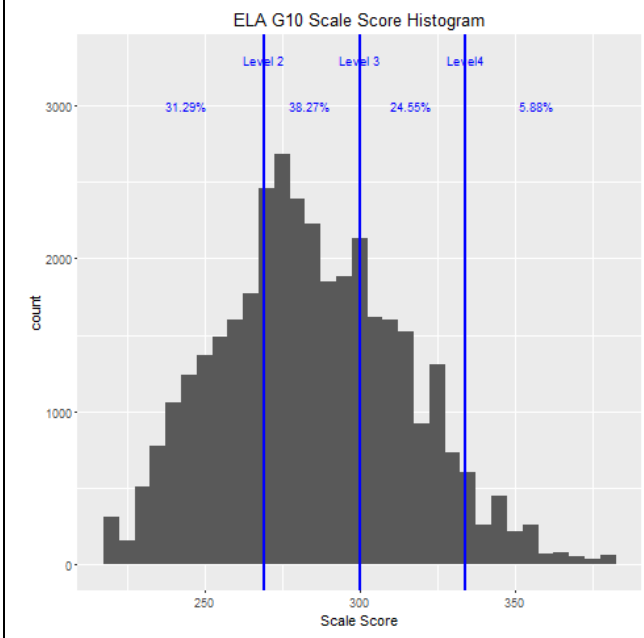
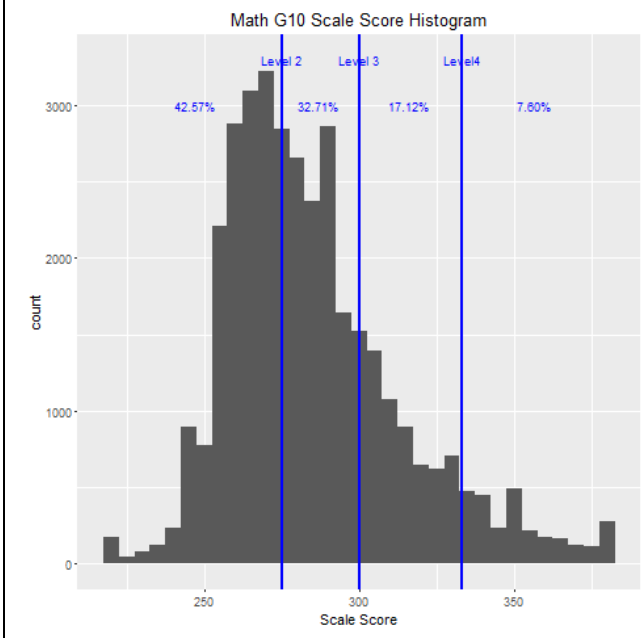
Grade	Path	Stage 1	Stage 2	*N*	%	Reliability
3	1	Medium	Easy	29,968	78.1%	0.95
3	2	Medium	Hard	8,424	21.9%	0.91
4	1	Medium	Easy	29,909	77.8%	0.95
4	2	Medium	Hard	8,559	22.3%	0.92
5	1	Medium	Easy	24,625	65.5%	0.95
5	2	Medium	Hard	12,948	34.5%	0.94
6	1	Medium	Easy	24,617	66.8%	0.94
6	2	Medium	Hard	12,258	33.2%	0.92
7	1	Medium	Easy	28,891	78.4%	0.93
7	2	Medium	Hard	7,947	21.6%	0.91
8	1	Medium	Easy	24,232	66.1%	0.93
8	2	Medium	Hard	12,443	33.9%	0.92
10	1	Medium	Easy	27,595	78.0%	0.94
10	2	Medium	Hard	7,789	22.0%	0.91

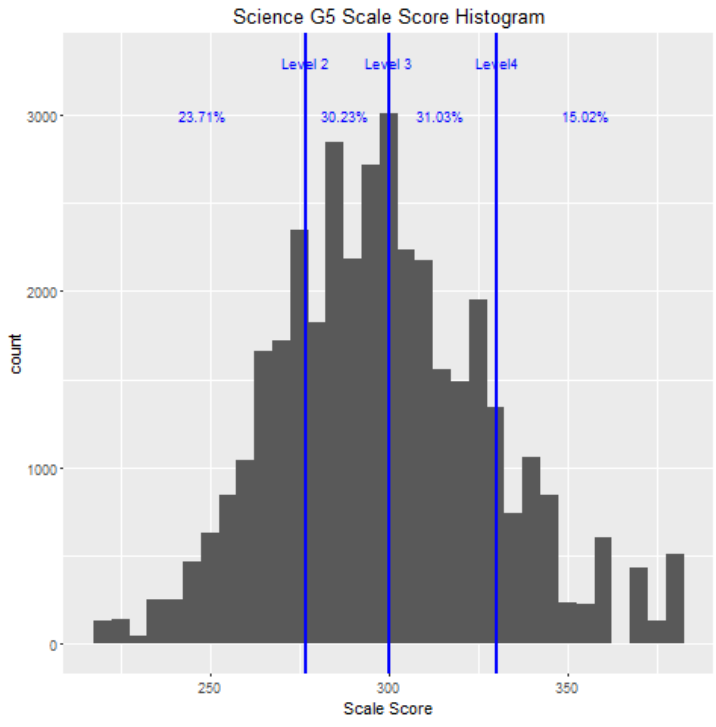
XIII. Appendix F: Scale Score Frequency Distribution

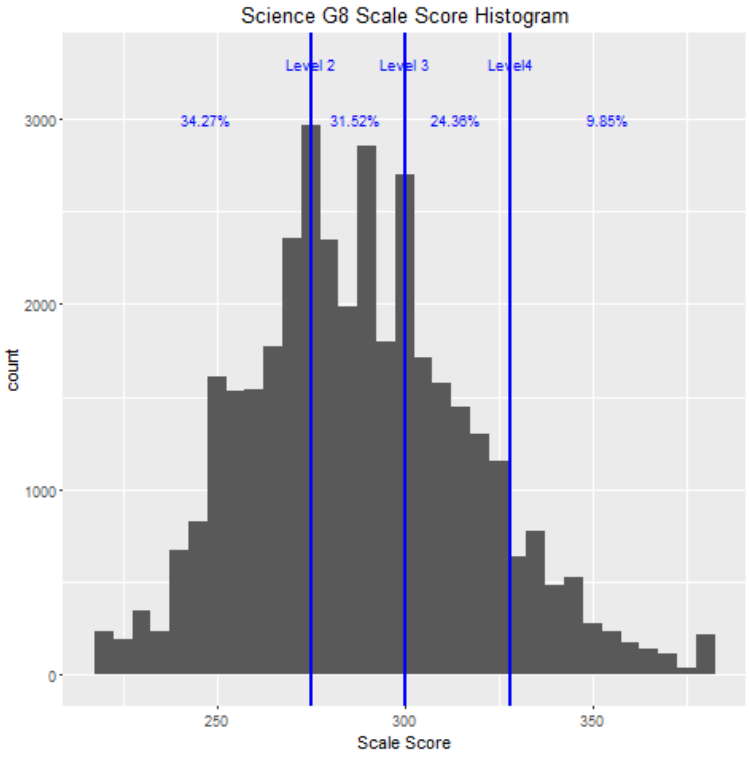
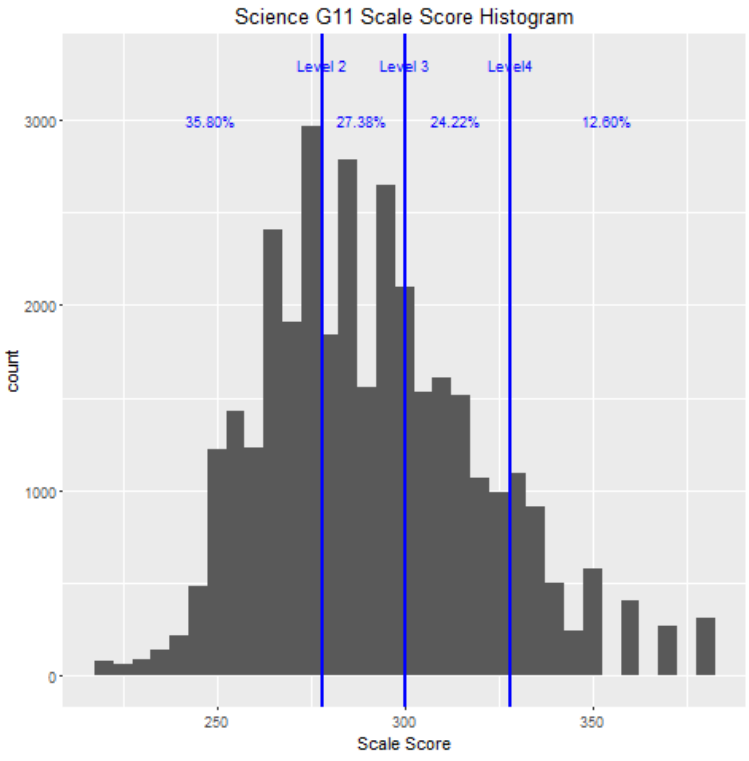
Grade	ELA Scale-Score Distribution	Math Scale-Score Distribution
3		
4		

5	 ELA G5 Scale Score Histogram Level 2 Level 3 Level4 24.08% 30.29% 28.87% 16.76% count Scale Score	 Math G5 Scale Score Histogram Level 2 Level 3 Level4 27.57% 22.45% 11.97% count Scale Score
6	 ELA G6 Scale Score Histogram Level 2 Level 3 Level4 32.33% 27.35% 34.97% 5.35% count Scale Score	 Math G6 Scale Score Histogram Level 2 Level 3 Level4 27.54% 23.50% 9.94% count Scale Score

7	 ELA G7 Scale Score Histogram Level 2 Level 3 Level4 33.11% 32.14% 25.59% 9.16% count Scale Score	 Math G7 Scale Score Histogram Level 2 Level 3 Level4 20.64% 49.02% 25.88% 4.47% count Scale Score
8	 ELA G8 Scale Score Histogram Level 2 Level 3 Level4 26.90% 45.01% 23.04% 5.05% count Scale Score	 Math G8 Scale Score Histogram Level 2 Level 3 Level4 40.00% 33.66% 19.85% 6.48% count Scale Score

10		

Grade	**Science Scale-Score Distribution**
5	

8	 Science G8 Scale Score Histogram Level 2 Level 3 Level4 34.27% 31.52% 24.36% 9.85% 3000 2000 1000 0 count 250 300 350 Scale Score
11	 Science G11 Scale Score Histogram Level 2 Level 3 Level4 35.80% 27.38% 24.22% 12.60% 3000 2000 1000 0 count 250 300 350 Scale Score

XIV. Appendix G: 2017 Kansas Assessment Survey Results

Table G-1. Demographics of Participants

Answer	%	Count
Total number responding		1259
District/Building Test Coordinator		238
District or Building Administrator		41
District Technology Coordinator, Curriculum Coordinator		12
Teacher		928
**Math teacher – 378*		
**ELA teacher – 409*		
**Science teacher – 177*		
**Classroom teacher – 165*		
*Special Education teacher – 89		
Other		40

*If teachers taught multiple subjects, they identified themselves as such. Individuals also identified themselves as a "Classroom teacher" in the "Other" category (290 teachers identified themselves in multiple categories).

Table G-2. Perceptions of Testing Length

For each statement please indicate if you agree, somewhat agree, disagree or are undecided.

	Agree	Somewhat Agree	Disagree	Undecided	Total
The length of the summative testing window was appropriate. (March 14 - April 28)	72.65% (906)	16.76% (209)	8.58% (107)	2.00% (25)	1247
The length of the ELA test was appropriate for students.	52.23% (620)	23.25% (276)	12.38% (147)	12.13% (144)	1187
The length of the Math test was appropriate for students.	51.14% (607)	22.16% (263)	14.32% (170)	12.38% (147)	1187
The length of the science test was appropriate for students.	42.79% (463)	14.60% (158)	4.81% (52)	37.80% (409)	1082

Table G-3. Rating of the Usefulness of Each Resource

Please rate the usefulness of each resource.

Resource	Extremely useful	Useful	Somewhat useful	Not useful	I did not use this resource.	Total
www.ksassessment.org website	7.35% 91	37.48% 464	18.17% 225	4.68% 58	32.31% 400	1238
KSDE list serve communications	9.65% 119	20.19% 249	12.08% 149	3.49% 43	54.58% 673	1233
District Test Coordinator training	11.35% 140	32.85% 405	14.76% 182	3.81% 47	37.23% 459	1233
KITE Service Desk	8.68% 107	15.33% 189	9.89% 122	3.24% 40	62.85% 775	1233
Kansas Examiner's Manual	16.28% 202	49.07% 609	21.92% 272	2.66% 33	10.07% 125	1241
Interactive demos (practice tests)	13.06% 162	29.52% 366	18.63% 231	6.29% 78	32.50% 403	1240
KITE Educator Portal	17.26% 215	36.12% 450	17.17% 214	4.57% 57	24.88% 310	1246
HELP tab inside the Educator Portal	4.29% 53	18.95% 234	12.96% 160	5.99% 74	57.81% 714	1235

Table G-4. Overall Feeling about Examiner's Manual, KITE, Educator Portal, and Assessments

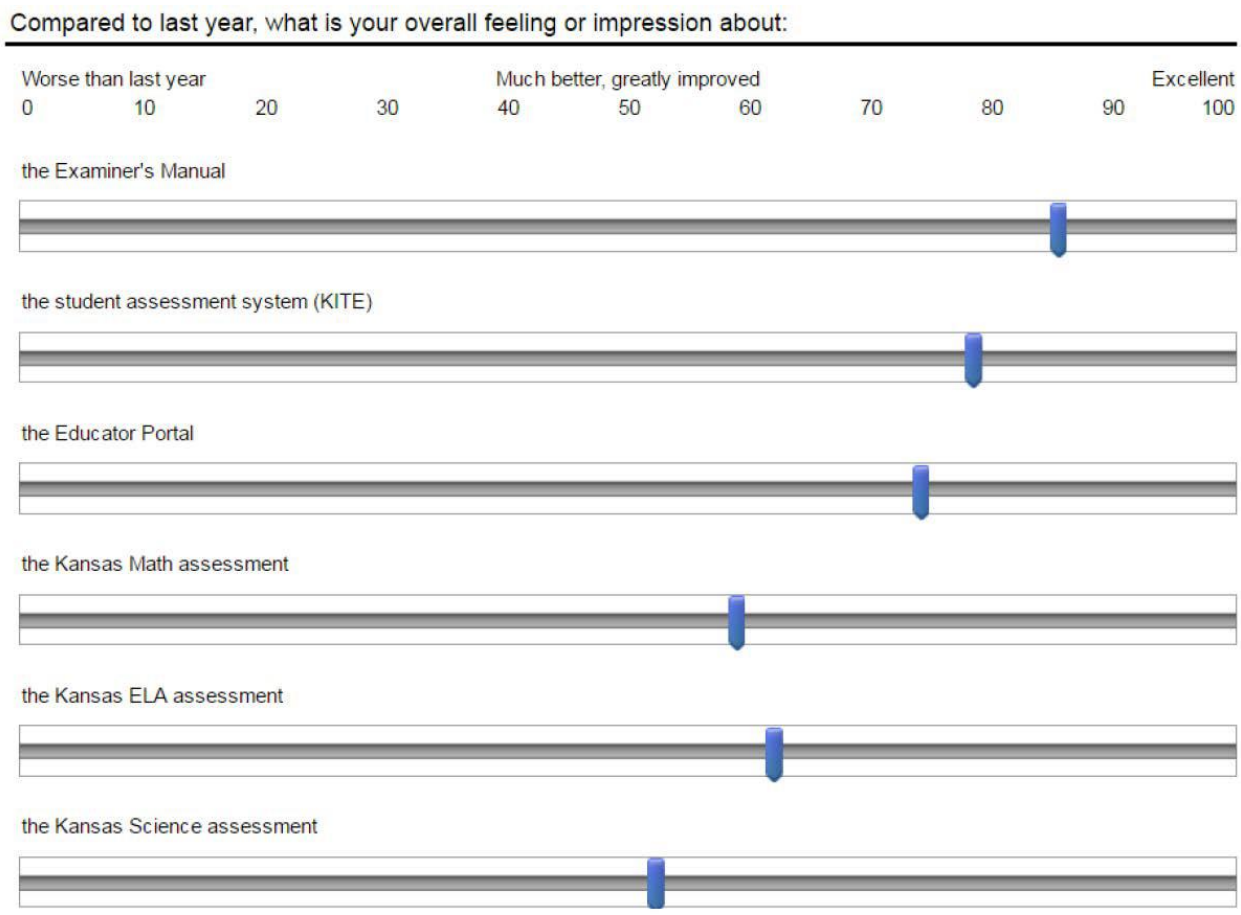


Table G-5.1. Overall Experience with KITE Service Desk

KITE Service Desk

Did you personally contact the KITE Service Desk this year?

	%	Count
By phone	13.30%	167
By email	10.91%	137
I did not contact the KITE Service Desk	83.44%	1048
Total	100%	1256

Approximately how many times did you contact the KITE Service Desk this year?

Answer	%	Count
1-5	65.13%	127
6-10	19.49%	38
11-15	7.18%	14
16 or more times	8.21%	16
Total	100%	195

Table G-5.2.1. Rating of Timeliness of Response by KITE Service Desk Staff

Rate your overall response to:
Timeliness of response by Service Desk staff

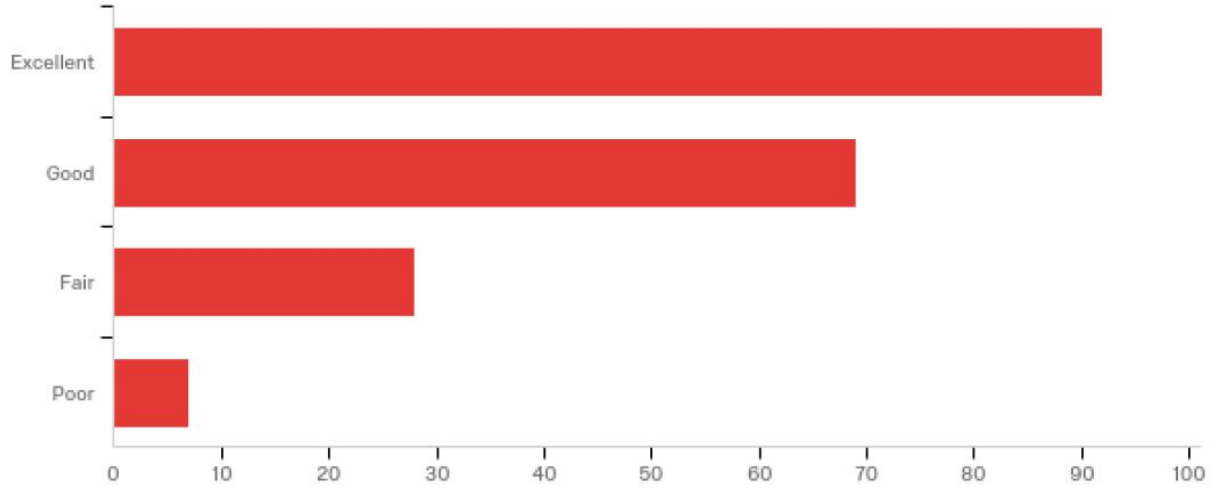


Table G-5.2.2. Rating of Timeliness of Response by KITE Service Desk Staff – Year-to-Year Comparison

Timeliness of response by Service Desk Staff – year-to-year comparison

	2017		2016	
Response	%	Count	%	Count
Excellent	46.94%	92	11.81%	15
Good	35.20%	69	36.22%	46
Fair	14.29%	28	33.86%	43
Poor	3.57%	7	18.11%	23
Total	100%	196	100%	127

Table G-5.3.1. Rating of Professionalism of KITE Service Desk Staff

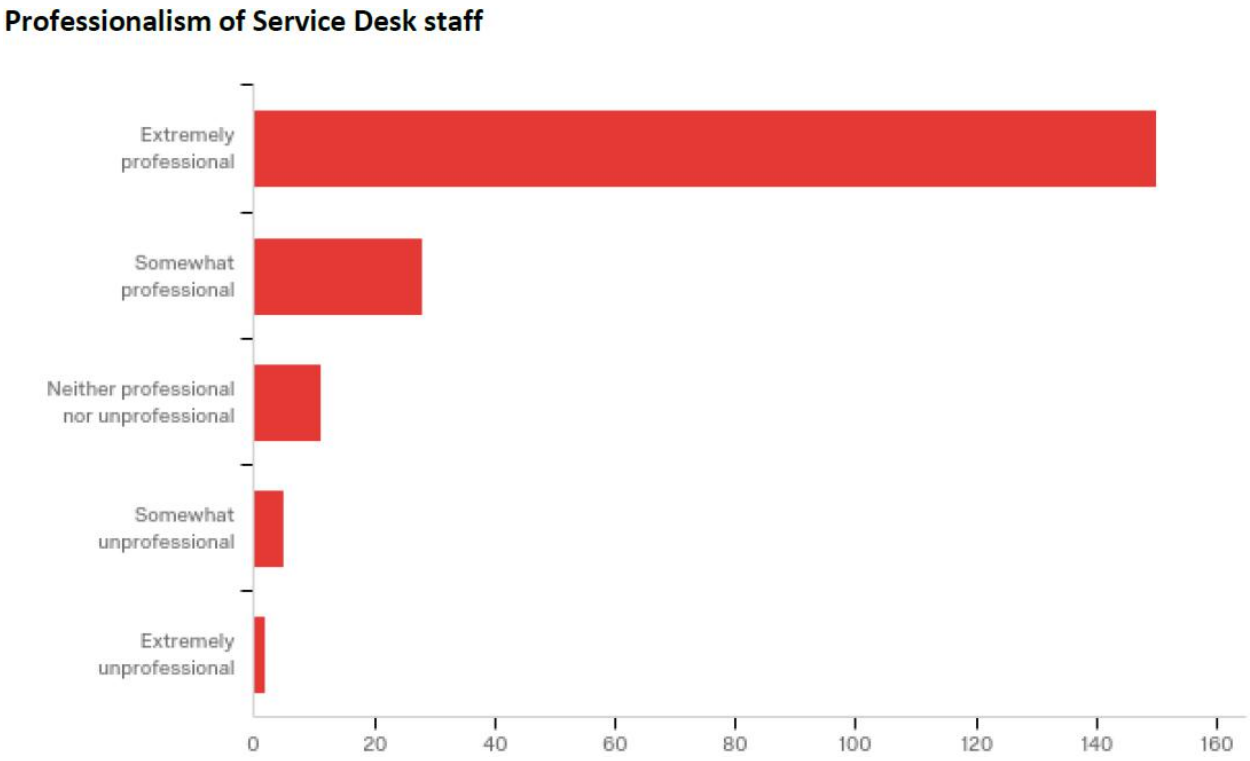


Table G-5.3.2. Rating of Professionalism of KITE Service Desk Staff – Year-to-Year Comparison

Year to Year Comparison	2017		2016	
Response	%	Count	%	Count
Extremely professional	76.53%	150	0%	0
Somewhat professional	14.29%	28	48.82%	62
Neither professional nor unprofessional	5.61%	11	36.22%	46
Somewhat unprofessional	2.55%	5	11.81%	15
Extremely unprofessional	1.02%	2	3.14%	4
Total	100%	196	100%	127

Note: Column percentages may not total to 100% due to rounding.

Table G-5.4.1. Rating of Knowledge of KITE Service Desk Staff

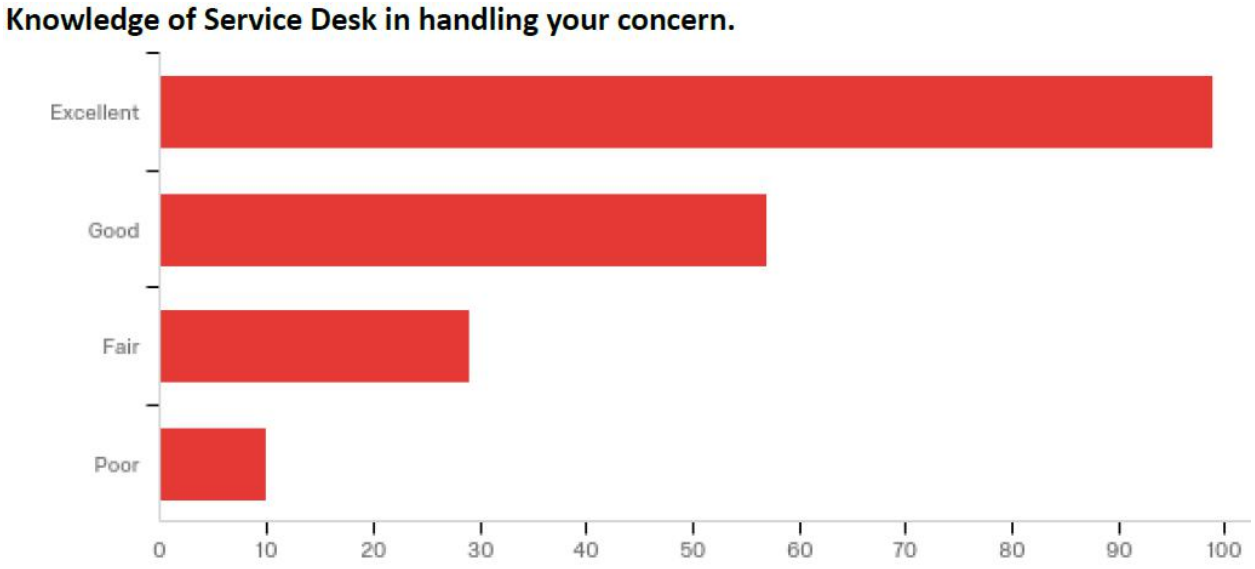


Table G-5.4.2. Rating of Knowledge of KITE Service Desk Staff – Year-to-Year Comparison

	2017		2016	
Response	%	Count	%	Count
Excellent	50.77%	99	24.80%	31
Good	29.23%	57	37.60%	47
Fair	14.87%	29	27.20%	34
Poor	5.13%	10	10.40%	13
Total	100%	195	100	125

XV. Appendix H: Science Performance Level Descriptors (PLDs)

Grade 5 Science		
Level 2	Level 3	Level 4
A student at Level 2 shows a basic ability to understand and use the science skills and knowledge needed for college and career readiness.	A student at Level 3 shows an effective ability to understand and use the science skills and knowledge needed for college and career readiness.	A student at Level 4 shows an excellent ability to understand and use the science skills and knowledge needed for college and career readiness.
Students who score at Level 2 can typically • use a model to describe that matter is made of particles too small to be seen, • state whether a new substance is produced by mixing substances, • identify evidence that plants primarily need air and water to grow, • describe the ways in which the four Earth spheres interact, • describe observable daily patterns of shadows and seasonal changes in the night sky, • describe a possible solution to an engineering problem.	Students who score at Level 3 can typically • develop a model to describe that matter is made of particles too small to be seen, • investigate whether the mixing of substances produces a new substance, • use evidence to support an argument that plants primarily need air and water to grow, • develop a model to describe the ways in which the four Earth spheres interact, • graph data to reveal observable daily patterns of shadows and seasonal changes in the night sky, and • generate and compare multiple possible solutions to an engineering design problem.	Students who score at Level 4 can typically • develop models to explain different types of matter made of particles too small to be seen, • investigate and provide evidence for whether the mixing of substances produces a new substance, • use evidence and models to support an argument that plants primarily need air and water to grow, • develop models to describe multiple ways in which the four Earth spheres interact, • graph data to explain observable daily patterns of shadows and seasonal changes in the night sky, and • use several sources to generate and compare multiple possible solutions to an engineering problem.

Grade 8 Science		
Level 2	Level 3	Level 4
A student at Level 2 shows a basic ability to understand and use the science skills and knowledge needed for college and career readiness.	A student at Level 3 shows an effective ability to understand and use the science skills and knowledge needed for college and career readiness.	A student at Level 4 shows an excellent ability to understand and use the science skills and knowledge needed for college and career readiness.
Students who score at Level 2 can typically • describe that mass is conserved in a chemical reaction, • describe the relationships of kinetic energy to mass and speed of objects, • explain how photosynthesis moves matter and energy through organisms in cycles, • identify information how humans influence inheritance of traits in organisms, • describe human impacts on the environment, • describe evidence of past tectonic-plate motions, and • explain how to improve an engineering design through repeated testing.	Students who score at Level 3 can typically • develop a model to describe how mass is conserved in a chemical reaction, • construct and interpret data to describe the relationships of kinetic energy to mass and speed of objects, • use evidence to explain how photosynthesis moves matter and energy through organisms in cycles, • gather and synthesize information about how humans influence the inheritance of traits in organisms, • design a method to monitor or minimize human impacts on the environment, • analyze and interpret data that provide evidence of past tectonic-plate motions, and • develop a model to optimize an engineering design through repeated testing.	Students who score at Level 4 can typically • develop and use models to explain why mass is conserved in chemical reactions, • generate, collect, and interpret data to explain the relationships of kinetic energy to the mass and speed of objects, • collect and use evidence to explain how photosynthesis moves matter and energy through organisms in cycles, • gather, synthesize, and communicate information about how humans influence the inheritance of traits in organisms, • design and refine a method to monitor or minimize human impacts on the environment, • analyze and interpret data to develop models that provide evidence of past tectonic-plate motions, • develop a model and synthesize data to optimize an engineering design through repeated testing.
Grade 11 Science		

Grade 8 Science		
Level 2	Level 3	Level 4
Level 2	Level 3	Level 4
A student at Level 2 shows a basic ability to understand and use the science skills and knowledge needed for college and career readiness.	A student at Level 3 shows an effective ability to understand and use the science skills and knowledge needed for college and career readiness.	A student at Level 4 shows an excellent ability to understand and use the science skills and knowledge needed for college and career readiness.
Students who score at Level 2 can typically • use a mathematical representation to claim that momentum in a system is conserved, • identify the advantages of using digital information, • describe factors affecting biodiversity and ecosystem populations, • make a claim about the causes of genetic variation, • describe a solution that reduces human impacts on natural systems, • describe the carbon cycle within the four Earth spheres, and • identify the needs and trade-offs of an engineering design.	Students who score at Level 3 can typically • use a mathematical representation to support the claim that momentum in a system is conserved, • evaluate questions about the advantages of using digital information, • use mathematical representations to explain factors affecting biodiversity and ecosystem populations, • use evidence to make and defend a claim about the causes of inheritable genetic variation, • evaluate or refine a solution that is designed to reduce human impacts on natural systems, • develop a quantitative model to describe the carbon cycle within the four Earth spheres, and • evaluate a complex, real-world problem to prioritize the needs and trade-offs of an engineering design.	Students who score at Level 4 can typically • collect data to create a mathematical representation to support the claim that momentum in a system is conserved, • evaluate questions and data about the advantages of using digital information, • analyze data and use mathematical representations to explain factors affecting biodiversity and ecosystem populations, • use evidence and models to make and defend a claim about the causes of inheritable genetic variation, • develop and use a quantitative model to describe the carbon cycle within the four Earth spheres, • evaluate, refine, and communicate solutions that reduce human impacts on natural systems, and • optimize a solution to a complex, real-world problem using prioritized needs and trade-offs of an engineering design.

XVI. Appendix I: Science Standard-Setting Meeting Agenda

Kansas Assessment Program (KAP)
Standard Setting Meeting
June 20–21, 2017
Lawrence, KS

Tuesday, June 20, 2017

Time	Agenda Item / Activity	Key Participants
7:00 a.m. – 8:00 a.m.	**Breakfast & Check-In**	All participants
8:00 a.m. – 8:45 a.m.	**Welcome & Orientation** • Welcome and introductions • Training on standard setting	Dr. Laura Kramer
8:45 a.m. – 9:00 a.m.	**Break-Out Groups** • Panelists go to separate rooms for the grade-level groups	Grade **5** panelists Grade **8** panelists Grade **11** panelists
9:00 a.m. – 9:15 a.m.	**Introduction** • Introduction, reminders • Panelists finish the nondisclosure agreement and participant survey	-Room facilitators -Grades **5**, **8**, and **11** panelists
9:15 a.m. – 10:15 a.m.	**Take the Test** • Panelists take the operational test • Discuss items as needed	-Room facilitators -Grades **5**, **8, and 11** panelists
10:15 a.m. – 10:30 a.m.	**Break**	
10:30 a.m. – 11:45 a.m.	**Just-Barely Student Activity** • Introduce just-barely student activity • Outline attributes for just-barely student qualifications at each level	-Room facilitators -Grades **5**, **8, and 11** panelists
11:45 a.m. – 12:30 p.m.	**Lunch**	All participants
12:30 p.m. – 1:45 p.m.	**Just-Barely Student Activity** • Outline attributes for just-barely students qualifications at each level	-Room facilitators -Grades **5**, **8, and 11** panelists
1:45 p.m. – 2:00 p.m.	**Break** • Refreshments provided	
2:00 p.m. – 3:45 p.m.	**Practice** • Check out secure materials • Practice bookmarking	-Room facilitators -Grades **5**, **8, and 11** panelists
3:45 p.m. – 5:00 p.m.	**Item Knowledge & Skills** • Write item Knowledge & Skills for test items on OIB	-Room facilitators -Grades **5**, **8, and 11** panelists

XVII. Appendix J: Participant Survey

SCIENCE STANDARD-SETTING PARTICIPANT SURVEY

Grade Level: _____

Directions: Please circle or write your answers. Your responses will be aggregated in technical documents for this event. Your individual responses will not be reported or linked to you.

1. Gender: Female Male

2. Ethnicity:

White Hispanic or Latino Black or African America

Asian/Pacific Islander Native American or American Indian Other

3. Current Assignment: Classroom Teacher Educator (Non-Teacher) Other

a) If you are a Classroom Teacher, do you teach **special-needs** students? Yes No

b) If you are a Classroom Teacher, do you teach **EL** students? Yes No

c) If "Other" please provide additional information below (occupation, educational focus, etc.)

4. Are you familiar with the KCCRS for Science Standards? Yes No

5. Work Setting: Urban Suburban Rural

6. District Name: _____

7. How many years (total) have you been teaching? _____

8. Please list the grades and the number of years you taught science at each grade.

9. Please describe your professional development activities in science within the past two years:
(Please use the back if necessary.)

XVIII. Appendix K: Confidentiality Agreement and Statement of Original Work

CENTER FOR EDUCATIONAL TESTING AND EVALUATION

CONFIDENTIALITY AGREEMENT AND STATEMENT OF ORIGINAL WORK

Test security and student confidentiality are of utmost importance to the Center for Educational Testing and Evaluation (CETE). As an item writer and/or reviewer for one or more of CETE's testing programs, or as a strategic partner, external researcher, or program reviewer, you will have access to test questions and other materials that must be kept secure. These assessment materials, documents, data, and other information are privileged and confidential and may not be used, shared, discussed, or otherwise published with any person who has not signed this confidentiality agreement. Please treat all materials as confidential and proprietary.

You are asked not to reproduce any test questions or other materials, directly or indirectly, and not to disclose the content of these materials. CETE takes pride in ensuring equity for all test takers. Therefore, please do not put any examinee at an unfair advantage by sharing information regarding item content at any time but particularly with colleagues or in a public forum. Additionally, items, data, or related assessment materials may not be reproduced, copied, photographed, published, announced, or in any other way made public, including both traditional and social media.

Additionally, for item writers, test materials developed by you for CETE must be original work. Test materials developed by you for CETE cannot be previously published or under consideration for publication elsewhere; materials developed by you become the property of CETE and its assessment partner(s).

We are certain that you share our concern that all potential assessment materials be handled in a professional, secure, and confidential manner, and we ask for your adherence to these guidelines by signing below. This Confidentiality Agreement will be enforced by the KU Center for Research acting as fiscal agent on behalf of CETE. The Agreement will be construed under the laws of the state of Kansas and the venue for enforcement will be the Douglas County Kansas District Court.

_____ _____

Participant Name (please print) Date

_____

Participant Signature

XIX. Appendix L: Readiness Form

READINESS FORM

Subject: _____ Grade: _____

Panelist ID: _____ Binder #: _____

Table #: _____

COMPLETE QUESTIONS 1 AND/OR 2 BEFORE STARTING ROUND 1.

1. I have completed the orientation and training and understand the purpose of the standard setting event. I also clearly understand my role in this event and what I am being asked to do. I am ready to begin round 1.

YES _____ NO_____ Your Initials_____

If you answered ***NO***, please raise your hand and ask the facilitator for additional help or training.

NOTE: ANSWER QUESTION 2 ONLY IF YOU SAID "NO" TO QUESTION 1.

2. I have received additional help and training. I now clearly understand my role and the task that I am being asked to do. I am now ready to begin round 1.

YES _____ NO_____ Your Initials_____

XX. Appendix M: Evaluation Form

EVALUATION FORM

Your opinions will provide us with a basis for evaluating both the materials and the training. **DO NOT** put your name on this form. We want your opinions to remain anonymous.

I. OPENING SESSION, TRAINING, AND PRACTICE *Indicate the extent to which you agree or disagree with the following statements.*	STRONGLY DISAGREE	DISAGREE	SOMEWHAT DISAGREE	SOMEWHAT AGREE	AGREE	STRONGLY AGREE
a. The opening session provided adequate background about the assessment program.						
b. The opening session provided a clear understanding of the purpose of the meeting.						
c. The opening session provided an appropriate context for my role in the meeting.						
d. The opening session addressed many of my questions and concerns.						
e. The opening session was well organized.						
f. The opening session leaders clearly explained the procedures.						
g. Taking the test helped me understand the assessment.						
h. The description of the ordered item booklet was clear.						
i. The presentation of the concept of the just-barely student was helpful.						
k. The training and practice helped me understand my tasks.						
l. The practice activities were effective.						
m. After training, I was able to complete the bookmark placement form accurately.						
n. After training, I understood my role in the event.						
o. The training facilitators effectively answered my questions.						

II. Training Time *Indicate how well the training time provided matched your need for training in this process.*	TOO LITTLE	ABOUT RIGHT	TOO MUCH
a. The amount of time used for training			

Use this space for additional comments you wish to share regarding the quality of the opening session, training, and practice.

III. The Just-Barely Student Activity *Indicate the degree to which you understood the just-barely student activity at each cut-score boundary.*	NO UNDERSTANDING	SLIGHT UNDERSTANDING	MODERATE UNDERSTANDING	COMPLETE UNDERSTANDING
Level 1/Level 2				
Level 2/Level 3				
Level 3/Level 4				

IV. Influential Factors for Round 1 *Indicate how important each of the following elements were as you placed your bookmarks for* ***Round 1****.*	NOT IMPORTANT	SLIGHTLY IMPORTANT	MODERATELY IMPORTANT	VERY IMPORTANT	NOT APPLICABLE
a. Descriptions of Level 1, Level 2, Level 3, and Level 4					
b. Your perceptions of the just-barely student					
c. Your perceptions of the difficulty of the items					
d. Your experience with students at this grade level					

IV. Influential Factors for Round 1 *Indicate how important each of the following elements were as you placed your bookmarks for* ***Round 1****.*	NOT IMPORTANT	SLIGHTLY IMPORTANT	MODERATELY IMPORTANT	VERY IMPORTANT	NOT APPLICABLE
e. Your experience with the Kansas Science Standards					

V. Influential Factors for Round 2 *Indicate how important each of the following elements were as you placed your bookmarks for* ***Round 2****.*	NOT IMPORTANT	SLIGHTLY IMPORTANT	MODERATELY IMPORTANT	VERY IMPORTANT	NOT APPLICABLE
a. Descriptions of Level 1, Level 2, Level 3, and Level 4					
b. Your perceptions of the just-barely student					
c. Your perceptions of the difficulty of the items					
d. Your experience with students at this grade level					
e. Your experience with the Kansas Science Standards					
f. Your Round 1 bookmark placements					
g. The bookmark placements of the other panelists during Round 1					
h. Group discussions					

VI. Influential Factors for Round 3 *Indicate how important each of the following elements were as you placed your bookmarks for* ***Round 3****.*	NOT IMPORTANT	SLIGHTLY IMPORTANT	MODERATELY IMPORTANT	VERY IMPORTANT	NOT APPLICABLE
a. Descriptions of Level 1, Level 2, Level 3, and Level 4					
b. Your perceptions of the just-barely student					
c. Your perceptions of the difficulty of the items					
d. Your experience with students at this grade level					
e. Your experience with the Kansas Science Standards					
f. Your Round 2 bookmark placements					

VI. Influential Factors for Round 3 *Indicate how important each of the following elements were as you placed your bookmarks for **Round 3**.*	NOT IMPORTANT	SLIGHTLY IMPORTANT	MODERATELY IMPORTANT	VERY IMPORTANT	NOT APPLICABLE
g. The bookmark placements of the other panelists during prior rounds					
h. Group discussions					
i. Impact data (i.e., percent of students at each achievement level)					

VII. Performance Level Descriptors (PLDs) *Indicate the degree to which you understand the PLDs at each level.*	NO UNDERSTANDING	SLIGHT UNDERSTANDING	MODERATE UNDERSTANDING	COMPLETE UNDERSTANDING
Level 1				
Level 2				
Level 3				
Level 4				

VIII. Performance level descriptors and the Bookmark Method *Indicate how applicable you think the PLDs are at each level for the Bookmark Method.*	NOT APPLICABLE	SLIGHTLY APPLICABLE	MODERATELY APPLICABLE	SIGNFICANTLY APPLICABLE
Level 1				
Level 2				
Level 3				
Level 4				

IX. The Just-Barely Student Activity and the Bookmark Method *Indicate how applicable you think the just-barely student activity is to the Bookmark Method.*	NOT APPLICABLE	SLIGHTLY APPLICABLE	MODERATELY APPLICABLE	SIGNFICANTLY APPLICABLE
Level 1/Level 2				
Level 2/Level 3				

IX. The Just-Barely Student Activity and the Bookmark Method Indicate how applicable you think the just-barely student activity is to the Bookmark Method.	NOT APPLICABLE	SLIGHTLY APPLICABLE	MODERATELY APPLICABLE	SIGNFICANTLY APPLICABLE
Level 3/Level 4				

X. Bookmark Placement Activities Indicate the extent to which you agree or disagree with the following statements.	STRONGLY DISAGREE	DISAGREE	SOMEWHAT DISAGREE	SOMEWHAT AGREE	AGREE	STRONGLY AGREE
a. The bookmark placement form was easy to understand.						
b. The expectations for each round were made clear.						
c. I made my ratings independently.						
d. I understood the tasks I was to accomplish for each round.						
e. I had the right amount of time to complete the tasks during each round.						

XI. Group Discussion Indicate the extent to which you agree or disagree with the following statements.	STRONGLY DISAGREE	DISAGREE	SOMEWHAT DISAGREE	SOMEWHAT AGREE	AGREE	STRONGLY AGREE
a. The group discussions aided my understanding of the issues.						
b. The time provided for discussions was adequate.						
c. Everyone had equal opportunity to contribute ideas and opinions.						
d. The discussions about the just-barely student were helpful to me.						
e. The discussions after the first round of rating were helpful to me.						
f. The discussions after the second round of rating were helpful to me.						

XII. MATERIALS *Indicate how useful each of the following elements were during the standard-setting process.*	NOT USEFUL	SLIGHTLY USEFUL	MODERATELY USEFUL	VERY USEFUL	NOT APPLICABLE
a. Performance Level Descriptors					
b. Item map table					
c. Item dot plot					
d. Items (in the ordered item booklet)					
f. Impact data (i.e., percent of students at each achievement level)					

Use this space for additional comments you wish to share regarding the item-rating activities.

XIII. ROUND 1 AND ROUND 2 RESULTS *Indicate the extent to which you agree or disagree with the following statements.*	STRONGLY DISAGREE	DISAGREE	SOMEWHAT DISAGREE	SOMEWHAT AGREE	AGREE	STRONGLY AGREE
*a. The **Round 1** results (e.g., tables, graphs) were **clear**.*						
*b. The **Round 1** results (e.g., tables, graphs) were **useful**.*						
*c. The **Round 2** results (e.g., tables, graphs) were **clear**.*						

XIII. Round 1 and Round 2 Results *Indicate the extent to which you agree or disagree with the following statements.*	*Strongly Disagree*	*Disagree*	*Somewhat Disagree*	*Somewhat Agree*	*Agree*	*Strongly Agree*
*d. The **Round 2** results (e.g., tables, graphs) were **useful**.*						

XIV. Grade-Level Group Results for the Level 2 Cut Score *Indicate the extent to which you agree or disagree with the following statements.*	*Strongly Disagree*	*Disagree*	*Somewhat Disagree*	*Somewhat Agree*	*Agree*	*Strongly Agree*
*a. The impact result (i.e., percentage of students) for this achievement level is **reasonable**.*						
*b. The cut score for this achievement level is **appropriate** based on the PLDs and the just-barely student activities.*						
*c. The cut score for this achievement level is **defensible** due to panelists' adherence to procedures.*						

XV. Grade-Level Group Results for the Level 3 Cut Score *Indicate the extent to which you agree or disagree with the following statements.*	*Strongly Disagree*	*Disagree*	*Somewhat Disagree*	*Somewhat Agree*	*Agree*	*Strongly Agree*
*a. The impact result (i.e., percentage of students) for this achievement level is **reasonable**.*						
*b. The cut score for this achievement level is **appropriate** based on the PLDs and the just-barely student activities.*						
*c. The cut score for this achievement level is **defensible** due to panelists' adherence to procedures.*						

XVI. Grade-Level Group Results for the Level 4 Cut Score *Indicate the extent to which you agree or disagree with the following statements.*	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
a. *The impact result (i.e., percentage of students) for this achievement level is **reasonable**.*						
b. *The cut score for this achievement level is **appropriate** based on the PLDs and the just-barely student activities.*						
c. *The cut score for this achievement level is **defensible** due to panelists' adherence to procedures.*						

Use this space for additional comments you wish to share regarding the results.

XVII. Agenda *Indicate how successful you believe each task or event was in the standard-setting process.*	Not Successful	Slightly Successful	Moderately Successful	Very Successful
a. *Opening session*				
b. *Taking an operational item set*				
d. *Discussions about the just-barely student*				
e. *Practice activities*				

XVII. AGENDA *Indicate how successful you believe each task or event was in the standard-setting process.*	NOT SUCCESSFUL	SLIGHTLY SUCCESSFUL	MODERATELY SUCCESSFUL	VERY SUCCESSFUL
f. Discussions of items' knowledge and skills				
g. Discussions after Round 1 bookmark placement				
h. Discussions after Round 2 bookmark placement				

XVIII. AAI STAFF *Indicate how helpful you felt each staff member was during the standard-setting process.*	NOT HELPFUL	SLIGHTLY HELPFUL	MODERATELY HELPFUL	VERY HELPFUL
a. Psychometric lead (trainer)				
b. Room facilitator				
c. Content specialists				
d. Other staff (please specify here): _____				
_____				
_____				

Use this space for additional comments you wish to share overall.

XXI. Appendix N: Articulation Session Evaluation Form

Evaluation of the Articulation Session. How clear was each of the following processes/materials?

Category	Description / Materials	Not clear	Slightly clear	Moderately clear	Very clear
Standard-Setting Processes	The process by which cut-score recommendations were made (i.e., the Bookmark Method)				
	The process by which 'just-barely' students were defined				
Articulation Meeting Description	The purpose of the articulation meeting				
	Your role as a panelist in the articulation meeting				
Data	The recommended cut-score results from the standard setting meeting				
	The information in the impact data tables				
	The "smoothed" results after articulation.				
Final Cut Scores	The process of making final cut-score recommendations				
	Your understanding that cut scores from this meeting will be used to inform KSDE's final recommendation to the State Board.				
Please provide any additional comments you have about the articulation session.					

XXII. Appendix O: Score Reports

Student Report

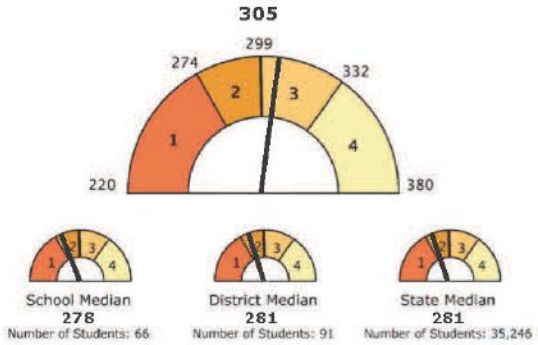
STUDENT REPORT: Matthews, Zoe
GRADE: 5 English Language Arts / STATE ID: 000000000
SCHOOL: Marysville Elementary
DISTRICT: Lorem ipsum Lorem ipsum orem / #D0511

2016—2017
KANSAS ASSESSMENT PROGRAM

The KAP assessments measure students' understanding of the Kansas College and Career Ready Standards at each grade. The English language arts assessment asks students to read and answer questions about literary passages, informational texts, and writing samples. Students demonstrate their knowledge and skills related to reading and writing by selecting the right answer and sorting, matching, labeling, and ordering information.

English Language Arts Score: Level 3

Last year your student performed at Level 2.



Students who perform at this level can typically

- read and understand moderately complex grade-level read and understand moderately complex grade-level texts
- summarize themes
- identify implied or clear details to support an idea
- determine meanings of more difficult words and complex figurative language
- identify literary elements and text structures and their impact on meaning
- determine point of view or purpose
- revise or edit a text to use academic language and correct grammar, punctuation, and spelling
- organize a text using sequence and logic
- determine if information is relevant
- use strategies to elaborate on ideas and structure texts

Performance Level Descriptions

Level 1: A student at Level 1 shows a limited ability to understand and use the English language arts skills and knowledge needed for college and career readiness.

Level 2: A student at Level 2 shows a basic ability to understand and use the English language arts skills and knowledge needed for college and career readiness.

Level 3: A student at Level 3 shows an effective ability to understand and use the English language arts skills and knowledge needed for college and career readiness.

Level 4: A student at Level 4 shows an excellent ability to understand and use the English language arts skills and knowledge needed for college and career readiness.

For more details about how your student performed on specific types of test questions, see the back of this report. →

STUDENT REPORT | STUDENT: Matthews, Zoe | STATE ID: 000000000 | GRADE: 5 English Language Arts

Your student's performance

Exceeds | Meets | Below | Insufficient Data

OVERALL READING

In this area, your student performed better than students who received the minimum Level 3 score. The reading portion requires students to read and analyze literary and informational texts and answer questions related to main ideas, text structure, language use, word meanings, and making and supporting conclusions.

READING: Literary Texts

In this area, your student performed better than students who received the minimum Level 3 score. This portion requires students to answer questions based on literary texts (such as stories and poems).

READING: Informational Texts

In this area, your student performed below students who received the minimum Level 3 score. This portion requires students to answer questions based on informational texts (such as science articles and historical speeches).

READING: Making and Supporting Conclusions

In this area, your student performed as well as students who received the minimum Level 3 score. These questions require students to read literary and informational texts and then make conclusions and use details and evidence to support ideas.

READING: Main Idea

In this area, your student performed better than students who received the minimum Level 3 score. These questions require students to read literary and informational texts and then determine central ideas, key events, and topics and identify supporting details.

OVERALL WRITING

In this area, your student performed below students who received the minimum Level 3 score. The writing portion requires students to read short writing samples and answer questions related to revising, editing, vocabulary, and language use.

WRITING: Revising

In this area, your student performed below students who received the minimum Level 3 score. These questions require students to revise provided text by applying writing skills, including using specific story-telling strategies, revising text into a logical order, adding context and detail, and identifying words or phrases to strengthen the text.

WRITING: Editing

In this area, your student performed as well as students who received the minimum Level 3 score. These questions require students to clarify messages in a variety of texts by following grade-appropriate grammar, capitalization, punctuation, and spelling rules.

WRITING: Vocabulary and Language Use

In this area, your student performed better than students who received the minimum Level 3 score. These questions require students to revise texts by using accurate language and vocabulary that is appropriate to a text's purpose and audience.

Standard error of measurement for this report:
Student—8.0 | School—3.5 | District—3.2 | State—0.2
The standard error indicates how much a student's score might vary if the student took many equivalent versions of the test (tests with different items but covering the same knowledge and skills).

Additional Resources
For sample test questions, go to **ksassessments.org/interactive-demos**.
For information on the Kansas College and Career Ready Standards, visit **ksde.org**.
To learn about the Kansas Assessment Program, go to **ksassessments.org**.
To discover more about this score report, see the 2017 Parent Guide at **ksassessments.org/pg** and in Spanish at **ksassessments.org/pg-esp**.

Kansas state department of Education

©2017 The University of Kansas

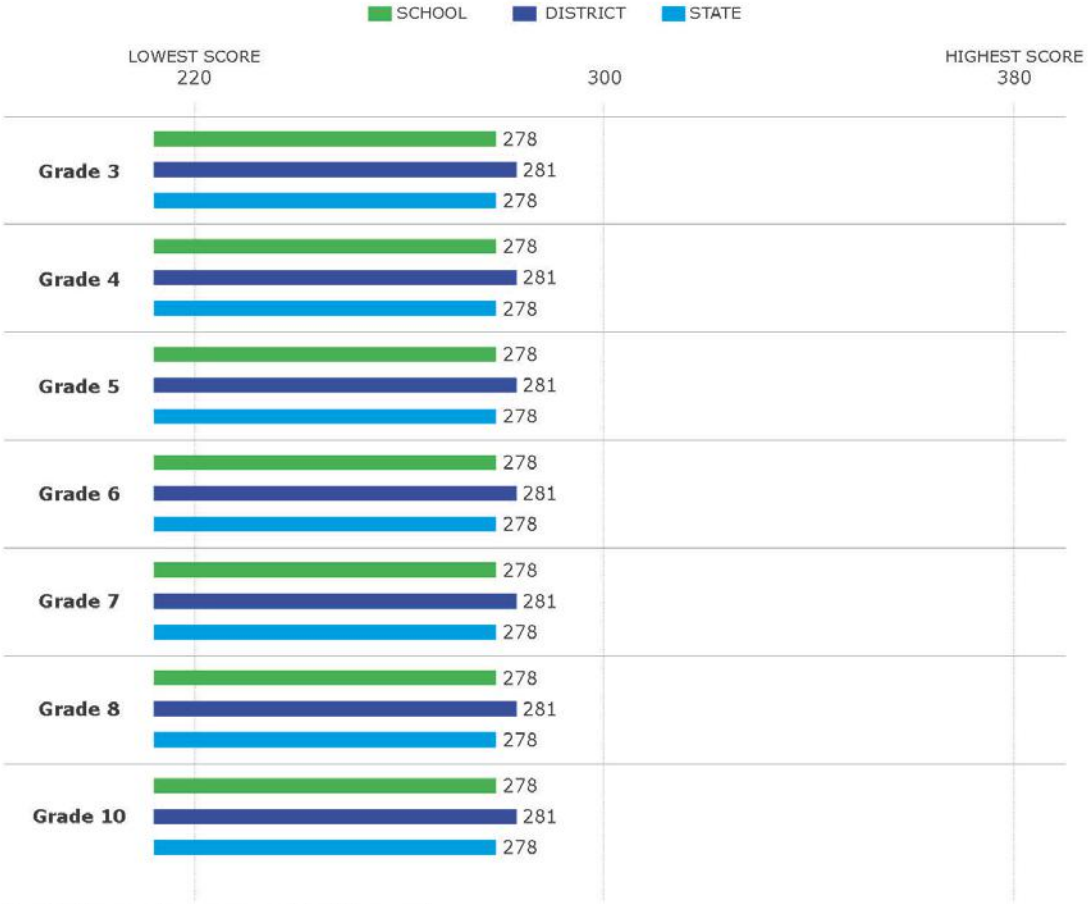
School Report

SCHOOL REPORT: Lorem ipsum orem / #D0511
SUBJECT: English Language Arts
DISTRICT: District Name / #

2016—2017
KANSAS ASSESSMENT PROGRAM

The KAP assessments measure students' understanding of the Kansas College and Career Ready Standards at each grade. The English language arts assessment asks students to read and answer questions about literary passages, informational texts, and writing samples. Students demonstrate their knowledge and skills related to reading and writing by selecting the right answer and sorting, matching, labeling, and ordering information.

Median School, District, and State Performance

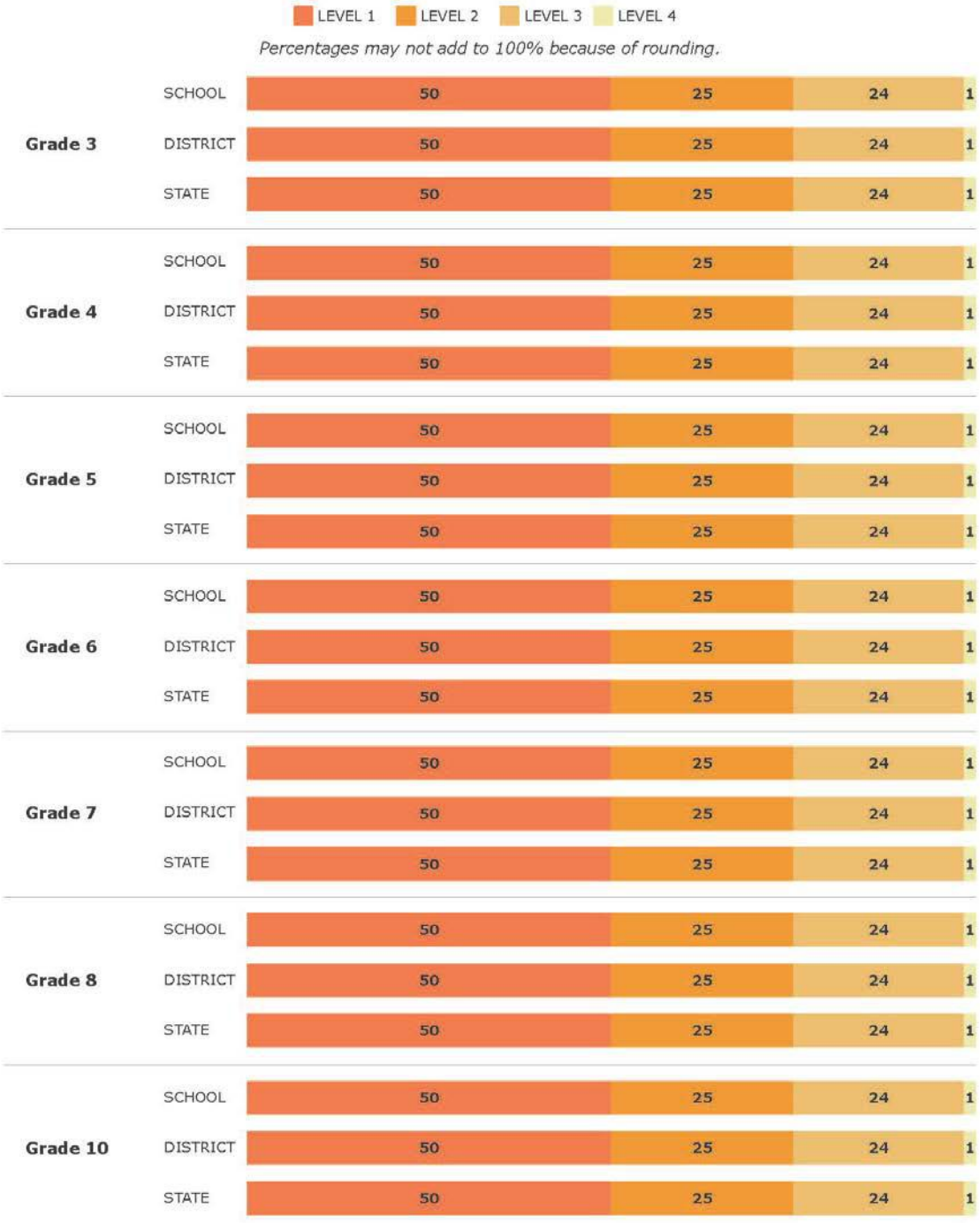


Standard error of measurement for this report:

Grade 3: School—3.5 | District—3.2 | State—0.2
Grade 4: School—3.5 | District—3.2 | State—0.2
Grade 5: School—3.5 | District—3.2 | State—0.2
Grade 6: School—3.5 | District—3.2 | State—0.2
Grade 7: School—3.5 | District—3.2 | State—0.2
Grade 8: School—3.5 | District—3.2 | State—0.2
Grade 10: School—3.5 | District—3.2 | State—0.2

The standard error indicates how much students' scores might vary if the students took many equivalent versions of the test (tests with different items but covering the same knowledge and skills).

Percentage of Students in Each Performance Level, by Grade



Your School's Performance

Exceeds | Meets | Below | Insufficient Data

Grade	3	4	5	6	7	8	10
OVERALL READING	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Literary Texts	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Informational Texts	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Making and Supporting Conclusions	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Main Idea	⊕	⊕	⊕	⊕	⊕	⊕	⊕
OVERALL WRITING	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Revising	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Editing	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Vocabulary and Language Use	⊕	⊕	⊕	⊕	⊕	⊕	⊕

OVERALL READING
The reading portion requires students to read and analyze literary and informational texts and answer questions related to main ideas, text structure, language use, word meanings, and making and supporting conclusions.

Literary Texts
This portion requires students to answer questions based on literary texts (such as stories and poems).

Informational Texts
This portion requires students to answer questions based on informational texts (such as science articles and historical speeches).

Making and Supporting Conclusions
These questions require students to read literary and informational texts and then make conclusions and use details and evidence to support ideas.

Main Idea
These questions require students to read literary and informational texts and then determine central ideas, key events, and topics and to identify supporting details.

OVERALL WRITING
The writing portion requires students to read short writing samples and answer questions related to revising, editing, vocabulary, and language use.

Revising
These questions require students to revise provided text by applying writing skills, including using specific story-telling strategies, revising text into a logical order, adding context and detail, and identifying words or phrases to strengthen the text.

Editing
These questions require students to clarify messages in a variety of texts by following grade-appropriate grammar, capitalization, punctuation, and spelling rules.

Vocabulary and Language Use
These questions require students to revise texts by using accurate language and vocabulary that is appropriate to a text's purpose and audience.

Your School's Performance

Exceeds
In this area, your students typically performed better than students who received the minimum Level 3 score.

Meets
In this area, your students typically performed as well as students who received the minimum Level 3 score.

Below
In this area, your students typically performed below students who received the minimum Level 3 score.

Insufficient Data
In this area, your students did not answer enough questions for accurate reporting.

Additional Resources

ACT Scoring
To get an idea of how your high school student may perform on the ACT based on this KAP score, refer to this chart. For more information, go to **ksassessments.org/act**.

Student's actual KAP grade 10 ELA score	Student's projected ACT reading score	Student's projected ACT English score
Level 1: 220–268	1–17	1–15
Level 2: 269–299	17–23	15–22
Level 3: 300–333	23–30	22–30
Level 4: 334–380	30–36	30–36

For sample test questions, go to **ksassessments.org/interactive-demos**.
For information on the Kansas College and Career Ready Standards, visit **ksde.org**.
To learn about the Kansas Assessment Program, go to **ksassessments.org**.
To discover more about this score report, see the 2017 Educator Guide at **ksassessments.org/eg**.

©2017 The University of Kansas

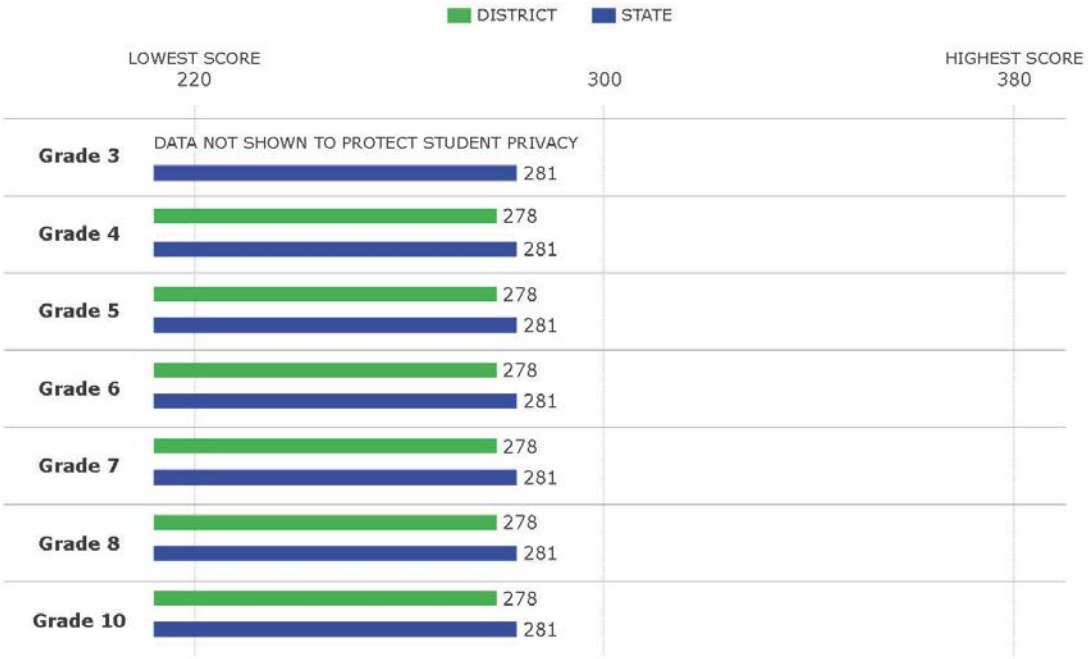
District Report

DISTRICT REPORT: Lorem ipsum orem / #D0511
SUBJECT: English Language Arts

2016–2017
KANSAS ASSESSMENT PROGRAM

The KAP assessments measure students' understanding of the Kansas College and Career Ready Standards at each grade. The English language arts assessment asks students to read and answer questions about literary passages, informational texts, and writing samples. Students demonstrate their knowledge and skills related to reading and writing by selecting the right answer and sorting, matching, labeling, and ordering information.

Median District and State Performance

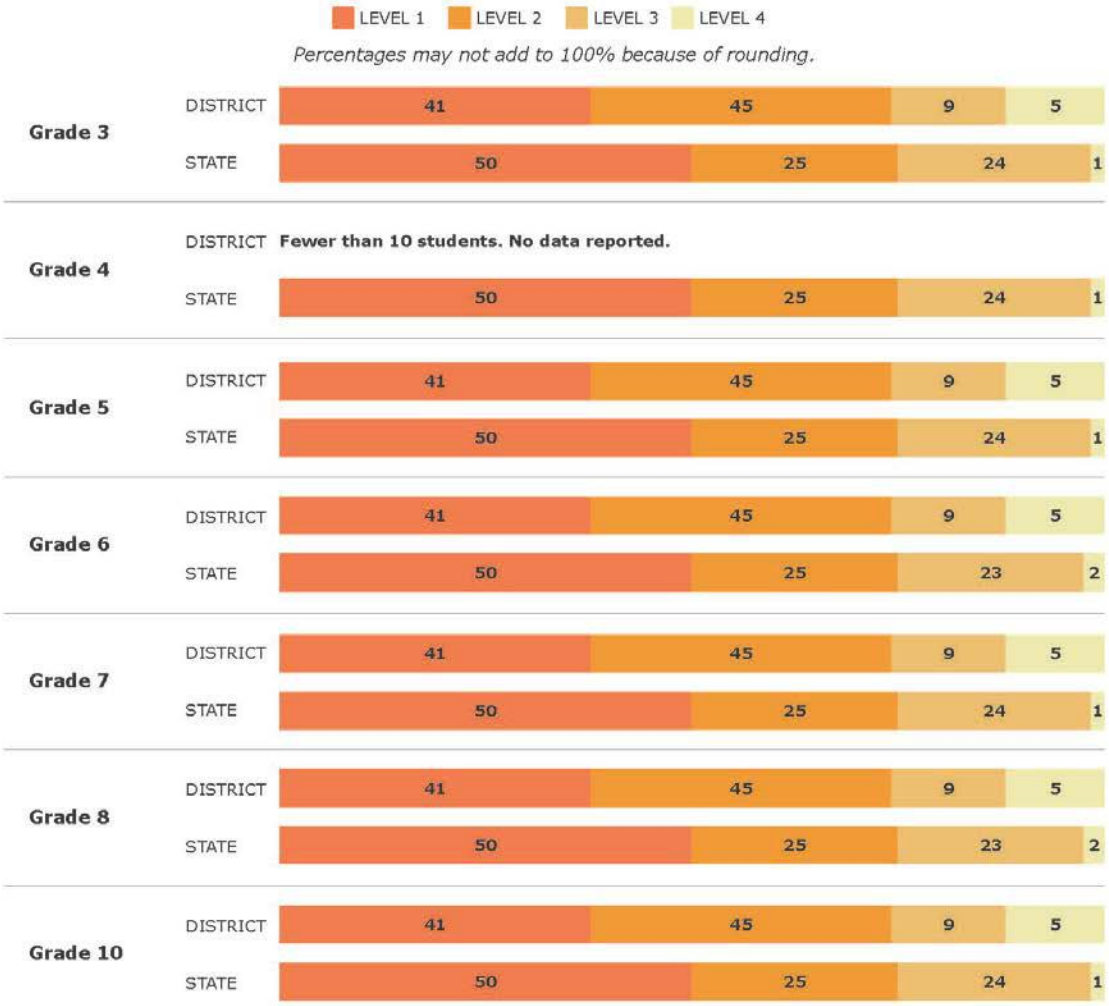


Standard error of measurement for this report:

Grade 3: District – 3.2 | State – 0.2
Grade 4: District – 3.2 | State – 0.2
Grade 5: District – 3.2 | State – 0.2
Grade 6: District – 3.2 | State – 0.2
Grade 7: District – 3.2 | State – 0.2
Grade 8: District – 3.2 | State – 0.2
Grade 10: District – 3.2 | State – 0.2

The standard error indicates how much students' scores might vary if the students took many equivalent versions of the test (tests with different items but covering the same knowledge and skills).

Percentage of Students in Each Performance Level, by Grade



DISTRICT REPORT | DISTRICT: XXXX | SUBJECT: English Language Arts

Your District's Performance

⊕ Exceeds ⊜ Meets ⊖ Below ⊠ Insufficient Data

Grade	3	4	5	6	7	8	10
OVERALL READING	⊜	⊜	⊕	⊕	⊜	⊕	⊜
Literary Texts	⊜	⊕	⊕	⊕	⊕	⊕	⊜
Informational Texts	⊕	⊕	⊠	⊕	⊕	⊠	⊖
Making and Supporting Conclusions	⊖	⊕	⊕	⊠	⊕	⊜	⊕
Main Idea	⊕	⊕	⊕	⊕	⊕	⊕	⊕
OVERALL WRITING	⊕	⊖	⊕	⊠	⊕	⊠	⊕
Revising	⊕	⊕	⊜	⊖	⊕	⊕	⊕
Editing	⊜	⊜	⊕	⊕	⊜	⊕	⊕
Vocabulary and Language Use	⊕	⊕	⊕	⊠	⊕	⊕	⊕

OVERALL READING
The reading portion requires students to read and analyze literary and informational texts and answer questions related to main ideas, text structure, language use, word meanings, and making and supporting conclusions.

Literary Texts
This portion requires students to answer questions based on literary texts (such as stories and poems).

Informational Texts
This portion requires students to answer questions based on informational texts (such as science articles and historical speeches).

Making and Supporting Conclusions
These questions require students to read literary and informational texts and then make conclusions and use details and evidence to support ideas.

Main Idea
These questions require students to read literary and informational texts and then determine central ideas, key events, and topics and to identify supporting details.

OVERALL WRITING
The writing portion requires students to read short writing samples and answer questions related to revising, editing, vocabulary, and language use.

Revising
These questions require students to revise provided text by applying writing skills, including using specific story-telling strategies, revising text into a logical order, adding context and detail, and identifying words or phrases to strengthen the text.

Editing
These questions require students to clarify messages in a variety of texts by following grade-appropriate grammar, capitalization, punctuation, and spelling rules.

Vocabulary and Language Use
These questions require students to revise texts by using accurate language and vocabulary that is appropriate to a text's purpose and audience.

Your District's Performance

Exceeds
In this area, your students typically performed better than students who received the minimum Level 3 score.

Meets
In this area, your students typically performed as well as students who received the minimum Level 3 score.

Below
In this area, your students typically performed below students who received the minimum Level 3 score.

Insufficient Data
In this area, your students did not answer enough questions for accurate reporting.

Additional Resources

ACT Scoring To get an idea of how your high school student may perform on the ACT based on this KAP score, refer to this chart. For more information, go to **ksassessments.org/act**.	Student's actual KAP grade 10 ELA score	Student's projected ACT reading score	Student's projected ACT English score
	Level 1: 220–268	1–17	1–15
	Level 2: 269–299	17–23	15–22
	Level 3: 300–333	23–30	22–30
	Level 4: 334–380	30–36	30–36

For sample test questions, go to **ksassessments.org/interactive-demos.**
For information on the Kansas College and Career Ready Standards, visit **ksde.org**.
To learn about the Kansas Assessment Program, go to **ksassessments.org**.
To discover more about this score report, see the 2017 Educator Guide at **ksassessments.org/eg**.

©2017 The University of Kansas

XXIII. Appendix P: Letters from the Commissioner of Education

KAP KANSAS ASSESSMENT PROGRAM

2017 Educator Guide

Understanding the Kansas Assessment Program Score Report

Dear Educators:

Thank you for your participation in the 2017 Kansas Assessment Program.

While assessments are an important tool that can help gauge a student's progress, we recognize they are just one of several measures to consider. Your use of classroom interaction, homework, assessments and other strategies throughout the year are equally important to the process of identifying learning and achievement levels.

The Kansas State Board of Education's vision for education — Kansas leads the world in the success of each student — reduces what many have considered an overemphasis on state assessments and increases the focus on the needs of the whole child. As we work toward this vision, you will see an increased focus on areas such as kindergarten readiness, Individual Plans of Study focused on career interest, high school graduation rates, postsecondary completion and social/emotional growth.

Kansas state department of Education

Assessments will continue to serve a role in helping to determine your students' academic readiness, but the State Board and the Kansas State Department of Education think it is time to minimize the assessment footprint on Kansas. We want the goals of each student — from the 5-year-old kindergarten student all the way to high school graduates considering a career, college or military — to be important.

Parents of 10th grade students will see a new ACT predictive measure added to their child's assessment report this year. The Center for Educational Testing and Evaluation has produced a report that correlates with or predicts a likely range of ACT scores based on how the student performed on the state assessment.

Kansas' teachers, students and parents are among the best in the nation, and we all share in the responsibility of making every child successful by achieving their desired future.

Thank you for all of your hard work and commitment to ensuring each student in Kansas is prepared for future success.

Sincerely,

Dr. Randy Watson
Kansas Commissioner of Education

KAP KANSAS ASSESSMENT PROGRAM

2017 Parent Guide

Understanding the Kansas Assessment Program Score Report

Dear Parents:

Thank you for supporting your child's participation in the 2017 Kansas Assessment Program.

While assessments are an important tool to help teachers, parents and students gauge a student's progress, it is essential to remember they are just one of several measures teachers consider. Your student's teachers use classroom interaction, homework, assessments and many other strategies throughout the year to identify learning and achievement levels.

The Kansas State Board of Education's vision for education — Kansas leads the world in the success of each student — reduces what many have considered an over-emphasis on state assessments, and increases focus on the needs of the whole child. As we work towards this vision, you will see schools focus on areas such as kindergarten readiness, Individual Plans of Study focused on career interest; high school graduation rates; postsecondary completion; and social/emotional growth.

Kansas state department of Education

Assessments will continue to serve a role in helping to inform your child's academic readiness, but the State Board and the Kansas State Department of Education (KSDE) believe it's time to minimize the assessment footprint on Kansas. We want the goals of each student — from the 5-year-old kindergarten student all the way to high school graduates focusing on career, college or military — to be important.

Parents of 10th grade students will see a new ACT predictive measure added to their child's assessment report this year. The Center for Educational Testing and Evaluation has produced a report that correlates with or predicts a likely range of ACT scores based on how the student performed on the state assessment.

As you review your child's report, please take the opportunity to contact your child's school, teacher or principal to have them explain these results to you in detail.

Kansas' students, teachers and parents are among the best in the nation, and we all share in the responsibility of making every child successful by achieving their desired future.

Thank you for being a positive part of your child's education, and thank you for your continued support of Kansas schools.

Sincerely,

Dr. Randy Watson
Kansas Commissioner of Education