Kansas Assessment Program
Technical Manual
2022

University of Kansas Achievement & Assessment Institute (AAI)

November 2022

KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

Table of Contents

KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

Table of Tables

Table of Figures

KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

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 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 three main purposes of KAP, as stated in the *Kansas Assessment Examiner's Manual 2021–2022,* are to

- measure specific claims related to the Kansas Standards in grades 3–8 and high school
- report individual student scores, along with each student's performance level
- 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. §72-5170 (2020). According to this statute, the State Board is mandated, in part, to

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

KAP provides the summative assessment in ELA, mathematics, and science for all students in grades 3–8 and high school, except students with significant cognitive disabilities, who are eligible for alternate assessments. The original KAP technical manual (i.e., the *2015 KAP Technical Manual*) was developed using 2014–2015 assessment data and published in April 2016. The technical manual was then updated each year, including technical-analysis results using that year's data and a description of new activities such as item development and standard setting. In the years with no changes to the assessment system or no new development, only the technical-analysis results were provided as an addendum. The following annual technical manuals can be found on the KAP website.

- *2015 KAP Technical Manual*
- *2016 KAP Technical Manual*
- *2017 KAP Technical Manual*
- *2018 KAP Technical Manual Addendum*
- *2019 KAP Technical Manual Addendum*
- *2020 KAP Technical Manual*
- *2021 KAP Technical Manual*

The current technical manual provides updates where applicable in ELA, mathematics, and science for the 2021–2022 school year, including a description of test forms used for the 2022 assessment, technical-analysis results using 2022 assessment data, and a summary of validity evidence to support the interpretation of test scores for intended test uses. For all subjects and grades, the test forms administered in 2020 were also administered in 2022, except grade-10

mathematics. The current technical manual describes the item and test development as well as standard setting for this new test.

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

For ELA and mathematics, the State Board adopted the Kansas Standards in 2010. The first administration of the operational KAP ELA and mathematics assessments aligned with the 2010 Kansas Standards occurred in 2015. More information about the 2010 Kansas Standards and KAP assessments can be found in the *2015 KAP Technical Manual* and the *2016 KAP Technical Manual.* In 2017, the State Board adopted the updated version of the 2010 Kansas Standards for ELA and mathematics. The planned 2020 [1] and current 2022 KAP ELA and mathematics assessments reflected the updated 2017 Kansas Standards.

The State Board adopted the Kansas Standards for Science in 2013. The first administration of the operational KAP science assessments aligned with the 2013 Kansas Standards occurred in 2017. In 2018, the Kansas science standards-review committee reviewed the 2013 Kansas science standards and concluded that no updates to the 2013 Kansas science standards were needed.

I.2. Coherent and Rigorous Academic Content Standards

Committees of Kansas educators and stakeholders provided input on the Kansas Standards. These standards supported the vision of the Kansas State Department of Education (KSDE): to lead the world in the success of each student (refer to the Kansas State Board of Education webpage). The standards help schools equip students with the academic, cognitive, metacognitive, technical, and employability skills required for postsecondary success, as well as the capacity to positively affect the world around them. The Kansas Standards are Kansas's coherent and rigorous academic content standards, which adhere to the State Board's mission. The mission of the State Board 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.

I.2.1. Goals of Kansas Standards

The 2017 Kansas Standards for ELA are built upon a foundation of common understandings, or practices, that provide a comprehensive view of broad goals for ELA and literacy instruction for each student across the state. The standards have five foundational practices:

1. Write, speak, read, and listen appropriately in all disciplines.
2. Seek out and work to understand diverse perspectives.
3. Use knowledge gained from literacy experiences to solve problems.
4. Create multimodal versions of texts for a range of purposes and audiences.

[1] The 2020 KAP spring administration was canceled because of the COVID-19 pandemic.

5. Self-regulate and monitor growth in writing, speaking, reading, and listening.

The 2017 Kansas Standards for Mathematics were created to define what students should understand and be able to do in their study of mathematics. Mathematical understanding is the ability to justify, in a way appropriate to the student's mathematical maturity, why a particular mathematical statement is true or where a mathematical rule comes from. The student who can explain the rule understands the mathematics and may have a better chance to succeed at a less familiar task. Mathematical understanding and procedural skills are equally important, and both are assessable using mathematical tasks of sufficient richness. With these standards for mathematics, Kansas leads the world in the success of each student in mathematics (refer to the Kansas State Board of Education Mathematics Standards webpage).

The closely align with the Next Generation Science Standards (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 provide an opportunity to improve student achievement in science; prepare students for college, career, and citizenship; and reflect a new vision for American science education.

I.2.2. Process and Timeline

The Kansas ELA standards-review committee met regularly to review and edit the previous 2010 standards document. After the committee completed its task, the updated standards were presented to the State Board in October 2017. Next, there was a window for public comments and review of the updated standards. The committee then presented the updated standards to the State Board in November 2017 for adoption, and the State Board adopted them later that month.

The previous Kansas mathematics standards were reviewed, written, and edited by the Kansas mathematics standards-writing-and-review committee between March 2016 and May 2017. Minutes of these standards-writing-and-review meetings were kept, explaining the decisions that were made (KSDE, 2017). The committee presented the updated standards to the State Board in July 2017 for adoption, followed by a window for public comments and review of the updated standards. In August 2017, the State Board approved the adoption (KSDE, 2017).

The 2013 Kansas Science Standards were reviewed in 2018 by the Kansas Science Standards-review committee. After review, the committee concluded that no updates to the 2013 Kansas Science Standards were needed. For the 2013 Kansas Science Standards, Kansas, as one the lead states in developing NGSS, had educators review the NGSS in 2013. These reviewers recommended the adoption of these standards to the State Board in 2013. The State Board adopted the NGSS as the 2013 Kansas Standards for Science in June 2013.

I.2.3. Standards-Review Committees

In an effort to ensure that educators from across the state had an opportunity to nominate either themselves or someone else to serve on the standards-review committees, information about the formation of the committees was distributed to the education community through email distribution lists, meetings, and the State Board. Nominations were collected via a registration

site that recorded the nominee's demographic information, committee group of interest, years of work experience, and highest level of education. KSDE staff ensured that the standards-review committees for ELA and science and the standards-writing-and-review committee for mathematics consisted of diverse genders, races, ethnicities, and teaching levels (K–12 and postsecondary) and that every state district was represented. Each committee also included an ad hoc group that consisted of representatives from various educational organizations, business communities, and KSDE, as well as legislators, parents, and other community members. Although the ad hoc group members participated in discussions during the standards-review process, they did not provide an official vote on the final product that was subsequently reviewed and adopted by the State Board (KSDE, 2017). As for the Kansas NGSS review committee in 2013, 60 members from across the state participated, comprising K–12 science educators, postsecondary science professors, and business and industry professionals.

I.3. Required Assessments and Intended Population

The KAP assessment measures student achievement in the subject areas of ELA, mathematics, and science. The subject areas and grades tested are ELA in grades 3–8 and 10, mathematics in grades 3–8 and 10, and science in grades 5, 8, and 11.

Kansas is committed to including 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 assessment, or the alternate assessment. Upon initial enrollment in Kansas schools, English learners are required to take the KAP mathematics and science tests. They are not required to take the ELA assessment but must take the Kansas English Language Proficiency Assessment. In their second year in Kansas schools, English learners are required to take all three KAP assessments.

Eligible students with significant cognitive disabilities, typically no more than 1% of Kansas students, take the Dynamic Learning Maps® (DLM®) alternate assessment for ELA, mathematics, and science. Other 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 (PNP). The PNP is a piece of information in a student's educational file that describes the accommodations provided to students during instruction. If an unapproved accommodation is used (e.g., reading aloud to a student on the KAP ELA test), the student test record is considered invalid. A detailed summary of accommodations for KAP can be found in Chapter V. Inclusion of All Students.

Exemptions from KAP assessments are granted to students who, during the testing window,

- move to a different school
- experience catastrophic illness or accident
- are serving long-term suspension
- are truant for more than two consecutive weeks
- are incarcerated in an adult facility
- are in a special detention center

II. Assessment System Operations

The development of any test requires many critical decisions regarding, for example, the content and cognitive complexity, the appropriate scope of that content for particular subject areas, and the number of items associated with each test. These decisions are not made in isolation but must be reasonable across all grade levels of the assessment. Together, these decisions guide the test-construction process and products.

II.1. Assessment Framework of the Assessed Grades

The assessment framework hierarchically categorizes the 2017 Kansas Standards for English language arts (ELA) and mathematics according to similar content. Those categories are classification, domain, and cluster. *Classification* is the largest category and consists of domains. *Domain* is the next category and consists of clusters. *Cluster* is the smallest category. A test item can be aligned to only one classification, one domain, and one cluster.

The ELA Standards are grouped by domain and cluster. ELA has two domains: reading and writing. Each grade's assessment measures all domains and clusters.

Mathematics Standards are grouped by classification, domain, and cluster. Mathematics has two classifications: skills and concepts, and strategic thinking and reasoning. Each grade's assessment measures all classifications but not all domains. The grade-10 mathematics assessment measures 11 domains, compared to three to five domains measured by other grades. Therefore, the domains within skills and concepts classification are grouped into conceptual categories for grade-10 mathematics.

Table II-1 and Table II-2 show the 2017 Kansas Standards assessment framework for ELA and mathematics.

Table II-1. English Language Arts Content Framework Across All Grades

Grade	Domain	Cluster
3–10	Reading	Information—key ideas and details
		Information—craft and structure
		Information—language in reading
		Information—integration of knowledge and ideas
		Literature—key ideas and details
		Literature—craft and structure
		Literature—language in reading
		Information—integration of knowledge and ideas
	Writing	Text types and purposes
		Language in writing

Table II-2. Mathematics Content Framework by Grade

Grade	Classification	Conceptual categories	Domain
3	Skills and concepts		Operations and algebraic thinking Numbers and operations with fractions Measurement and data Geometry
4	Skills and concepts		Operations and algebraic thinking Number and operations in base ten Numbers and operations with fractions Measurement and data
5	Skills and concepts		Number and operations in base ten Numbers and operations with fractions Measurement and data
6	Skills and concepts		Ratios and proportional relationships The number system Expressions and equations Geometry Statistics and probability
7	Skills and concepts		Ratios and proportional relationships The number system Expressions and equations Geometry Statistics and probability
8	Skills and concepts		Expressions and equations Functions Geometry
10	Skills and concepts	Number and quantity and algebra	The real number system Seeing structure in expressions Arithmetic with polynomials and rational expressions Reasoning with equations and inequalities
		Functions	Interpreting functions Building functions

Grade	Classification	Conceptual categories	Domain
		Geometry	Congruence
			Similarity, right, triangles, and trigonometry
			Expressing geometric properties with equations
		Statistics and probability	Interpreting categorical and quantitative data
3–10	Strategic thinking and reasoning		Strategic thinking and reasoning

The 2013 Kansas Standards for Science follow a different hierarchal structure. Science Standards are grouped by claims and targets. Targets are sublevels of claims. An item is aligned to only one claim and one target. Science has three claims: physical science, life science, and Earth and space science. In science, each grade's assessment assesses all claims, but not all targets. Table II-3 shows the 2013 Kansas Standards assessment framework for science by grade.

Table II-3. Science Content Framework by Grade

Grade	Claim	Target
5	Physical science	Structure and properties of matter Engineering design in physical science
	Life science	Matter and energy in organisms and ecosystems Engineering design in life science
	Earth and space science	Earth's systems Space systems Engineering design in Earth and space science
8	Physical science	Structure and properties of matter Chemical reactions Forces and interactions Energy Waves and electromagnetic radiation Engineering design in physical science
	Life science	Structure, function, and information processing Matter and energy in organisms and ecosystems Interdependent relationships in ecosystems Growth, development, and reproduction of organisms Natural selection and adaptations Engineering design in life science
	Earth and space science	Space systems History of the Earth Earth's systems Weather and climate Human impacts Engineering design in Earth and space science
11	Physical science	Structure and properties of matter Chemical reactions Forces and interactions Energy Waves and electromagnetic radiation Engineering design in physical science
	Life science	Structure and function Matter and energy in organisms and ecosystems Interdependent relationships in ecosystems Inheritance and variation of traits Natural selection and evolution Engineering design in life science

Earth and space science	Space systems History of the Earth Earth's systems Weather and climate Human sustainability Engineering design in Earth and space science

II.2. Test Design and Development

As described in Section I.2 Coherent and Rigorous Academic Content Standards, the 2017 Kansas Standards for ELA and Mathematics were adopted as updated versions of the 2010 Kansas Standards (except grade-10 mathematics); thus, the Achievement and Assessment Institute (AAI) worked with the Kansas State Department of Education (KSDE) to determine the content to be assessed by Kansas Assessment Program (KAP) tests for each subject area and grade so that 2022 KAP assessments supported continuity with the previous standards. For grade-10 mathematics, the 2017 Kansas Mathematics Standards added nine new domains under the skill and concept classification, removed seven domains under the skill and concept classification, and shifted performance level descriptors associated with standards. Given the extent of changes made to the grade-10 Mathematics Standards, a new test blueprint, an new assessment, and new achievement standards were warranted for the grade-10 mathematics test.

As described in Section I.2 Coherent and Rigorous Academic Content Standards, the 2013 Science Standards were reviewed in 2018 and carried forward with no changes or updates. The development leading to the original 2017 science-assessment administration, based on the 2013 standards, occurred over multiple years. No changes were made to the science test design following the 2018 review of the Science Standards. Table II-4 outlines the test-development timeline for ELA, mathematics, and science.

Table II-4. Development Timeline for KAP Assessments

Milestone	Date	Note
English language arts		
Adoption of 2010 Kansas Standards	October 2010	
First operational administration aligned to 2010 Kansas Standards	Spring 2015	
Standard setting	Summer 2015	
Item development	Summer 2017	
Adoption of 2017 Kansas Standards	November 2017	
Item realignment for 2017 Kansas Standards	2018 to 2020	
Field-testing items realigned to 2017 Kansas Standards	Spring 2019	Items are not included in scoring.

Milestone	Date	Note
Items aligned to 2017 Kansas Standards included in summative assessment	Spring 2022	All operational items are aligned to 2017 Kansas Standards.
Mathematics grades 3–8		
Adoption of 2010 Kansas Standards	October 2010	
First operational administration aligned to 2010 Kansas Standards	Spring 2015	
Standard setting	Summer 2015	
Adoption of 2017 Kansas Standards	August 2017	
Item realignment for 2017 Kansas Standards	2017 to 2020	
Field-testing items realigned to 2017 Kansas Standards	Spring 2019	Items are not included in scoring.
Items aligned to 2017 Kansas Standards included in summative assessment	Spring 2022	All operational items are aligned to 2017 Kansas Standards.
Mathematics grade 10		
Adoption of 2017 Kansas Standards	August 2017	
Finalize blueprint for an assessment aligned to 2017 Kansas Standards	January 2019	
Item development and realignment for 2017 Kansas Standards	2017–2021	
First operational administration aligned to 2017 Kansas Standards	Spring 2022	Operational field testing and all items are aligned to 2017 Kansas Standards.
Standard setting	June 2022	
Science		
Adoption of Kansas Standards	June 2013	
Kansas Standards item development	2015 to 2016	Determined annually
Census field testing	Spring 2016	Machine-scored items only
First operational administration aligned to Kansas Standards	Spring 2017	Machine-scored items only
Standard setting	Summer 2017	
Review of Kansas Standards	2018	No updates to the 2013 Kansas Standards

II.2.1. Test Blueprints

The blueprints were developed in collaboration among AAI content team, KSDE, and educators. Table II-5 summarizes the range of the proportion of items required for each domain in the test blueprints for ELA, for each classification in mathematics, and for each claim in science. The proportions did not vary across grades. Some ELA items were worth one point, while other items were worth two points. All mathematics and science items were worth one point.

Table II-5. Test Blueprint by Subject and Content Category for English Language Arts, Mathematics, and Science

Subject and content category	Items by category (%)
English language arts	
Reading	60–70
Writing	35–40
Mathematics	
Skills and concepts	75–88
Strategic thinking and reasoning	12–25
Science	
Physical science	27–33
Life science	34–40
Earth and space science	27–33

These blueprints for the current KAP assessments are also published in the assessment-development guides on the KAP website. The blueprints in the assessment-development guides also include the cognitive complexity level requirement and clusters or domains measured for each category besides item distribution. The grade-specific blueprints in the development guides are in Appendix A.

II.2.2. Test Design

In 2022, all three subjects used a fixed-form test design. Each subject had one operational form administered in two sessions. Each session offered several blocks of items that were the same but presented in different order to deter cheating. According to research, item orders do not affect item performance (Hohensinn et al., 2011; Li et al., 2012), so blocks with items in different order were still considered to be the same test form. Students were randomly assigned to one test form, and there was a designated test form for students who needed accommodations. Table II-6 shows the test design of the KAP assessment for each session by subject.

Table II-6. Fixed-Form Test Design of the 2022 KAP Assessment by Subject and Session

Subject	Grade	No. of items		
		Total	Session 1	Session 2
ELA	3–8, HS	47	22	25
Mathematics	3–8	55	25	30
Mathematics	HS	56	25	31
Science	5	35	18	17
Science	8, HS	40	20	20

Note. ELA = English language arts; HS = high school.

For all other subjects and grades except grade-10 mathematics, all assessment items have passed all item reviews, and test forms are operational forms. For grade-10 mathematics, 26 items that were previously developed items and were realigned have passed all item reviews, including data review, and were considered operational items. Another 30 operational items were needed to meet the blueprint for grade-10 mathematics. To select those 30 operational items, 38 items were operationally field tested in 2022. After operational field testing, 30 items were selected and combined with 26 items to construct the 56-item operational form.

II.2.3. Operational Test Construction

The 2022 test forms of all grades and subjects used the same procedures and guidelines that were used as in previous years:

- Items and passages were approved by KSDE prior to field testing and were reviewed by panels of external stakeholders for appropriateness and alignment.
- After field testing, items were reviewed for content and psychometric characteristics to rank items by preference of inclusion for assessments.
- Test sessions were assembled following the content specifications in the blueprint; items with the best psychometric characteristics were preferentially selected.
  - Items with negative or very low discrimination, or extremely low or high item difficulties were not selected.
  - Each test session included a wide range of item difficulties, and the average difficulty was of a moderate level.
- Test sessions were reviewed to eliminate item enemies (e.g., items that might clue answers to other items).
- Test sessions were reviewed and approved by psychometric staff for psychometric properties.
  - Ensuring test sessions included items with a wide range of item difficulties and with a moderate level of test difficulty.
  - Ensuring test sessions provided the maximum information about the medium theta level (i.e., theta between −0.5 and 0.5).
- For mathematics only, each test session begins with calculator-inactive items, followed by calculator-active items.

II.3. Item Development

Item development entailed 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 refresher training for continuing item writers, creating items, and reviewing and revising items. Item review was conducted in two phases: first, when items were created, and next, after items were field tested. In the first phase, both AAI content experts and trained, external item reviewers reviewed items.

Before appearing on any assessment, items were reviewed by content reviewers, fairness reviewers, and KSDE staff. The AAI content team used item-review feedback to revise test items as needed. Items were then prepared for field testing according to test specifications and established guidelines for both general and accommodated presentations. After field testing, AAI content experts and psychometricians analyzed the item and test data.

The next sections describe item development for the 2022 KAP assessments. ELA, mathematics, and science item development occurred from 2013 to 2017. ELA and mathematics items were first written to align to 2010 Kansas Standards and then realigned to the updated 2017 standards. New items were developed to align to the new grade-10 Mathematics Standards from 2018 to 2022. These items were developed to construct the 2022 grade-10 mathematics operational test form with previously realigned grade-10 mathematics items. However, item-review procedures were consistent across all subject areas and grades.

II.3.1. English Language Arts Passage Selection and Review

For ELA, the process starts with identifying appropriate public-domain works or commissioning passages as work-for-hire. AAI's content 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 AAI's content team uses several resources, both qualitative and quantitative, to analyze text complexity and guide grade placement. Assessment passages include commissioned, permissioned, and public-domain readings. Passages from all sources undergo multiple rounds of internal and external review. The *2017 KAP Technical Manual* provides more information about the passage-review process.

After passage review, AAI 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 Writing

ELA items were written internally at AAI in three item-development activities: one in 2012, one in 2013 and one in 2017. Mathematics items were written internally at AAI in two item-development activities: one in 2012 and one in 2013. Science items were written internally at AAI in a two-year span: from spring to winter in 2015, and from fall to winter in 2016. All item writers completed item-writing training and had content expertise in the relevant subject areas.

II.3.2.1. Item Writers

Item writers included full-time employees of AAI and graduate research assistants (GRAs) from the University of Kansas. GRAs were recruited and hired based on their training in a given subject, prior item-writing or test-development experience, or previous teaching experience. As ELA, mathematics, and science tests cover a wide range of knowledge and skills that incorporate diverse, real-life topics as item contexts, GRAs who wrote items for the assessments came from diverse academic backgrounds, including biology, classical languages, computer science, curriculum and teaching, economics, educational psychology, ELA, mathematics, physics, premedical, and social welfare.

II.3.2.2. Item-Writing Training

Before writing items for the KAP assessment, item writers trained in the use of KAP subject-area item specifications for writing and reviewing of items. All item writers received training in the following topics:

- Kansas Standards
- alignment
- bias and sensitivity
- differentiation between cognitive complexity and difficulty
- evidence-centered design
- item types
- principles of Universal Design for Learning (UDL) and accessibility
- validity and reliability
- item-writing best practices

Besides learning fundamental principles of item writing, item writers also received training in item review so they could objectively evaluate their own products as well as others' items. Key points of these writing and reviewing guidelines are described below.

- General guidelines
  - Write items that have clearly correct answer choices, 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.
- 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.

  - Align items to the cognitive complexity of content standards.
  - Write items to a variety of difficulty levels.
- 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.
- Structure guidelines
  - Avoid complex-format items.
  - Write items in the form of a question.
  - Avoid window-dressing items (e.g., excessive verbiage).
- 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.
  - Ensure that question asked by the item is as close to at the end of stem as possible.
  - Minimize the use of qualifying words (e.g., “best,” “most likely”).
- Answer-choice development guidelines
  - Order answer choices logically.
  - Create independent answer choices that do not overlap.
  - Write answer choices that are of roughly 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.
  - Ensure that answer keys should be roughly equally distributed among options for all items developed.
- Accessibility guidelines:
  - Consider the access needs of special populations and how 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.
- 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 included extensive practice. During practice, participants first discussed the depth of knowledge (DOK) framework (Webb, 1997) for specific standards, examined practice items for alignment to content standards, and determined whether example practice items were written to the appropriate difficulty level. Participants also practiced writing items and received feedback from AAI staff.

II.3.2.3. Item-Writing Process

Because of the research needed to ensure the descriptive information included in the item was technically correct, initial item writing ranged from a few hours to a few days. The item writer matched the item to the metadata requirements; the item writer also ensured that the item followed the rules of item writing, the content was correct and any surrounding context was accurate, the language was appropriate for the grade being tested, and then verified the correct answers.

The item writer sent each completed item to a fellow item writer for review. They discussed the item, the alignment to the standards, and the cognitive complexity demands and then revised items as needed. The items were then passed to a content specialist or test-development assistant for further review.

Following the content specialists' review, the item was sent to the editing team. If graphics were needed, a content specialist provided instructions to the graphic artist regarding the rendering of the stimulus, then confirmed that the completed graphic met the intended function within the item. When the editors finished editing the items, the content specialists reviewed the items before passing the set to the content lead and psychometricians for adherence to item-writing best practices.

The content lead approved the item (and graphics if needed), made their own edits, or sent the item back to the item writer, content specialist, or graphic artist for revision. Items were then reviewed by the content lead for adherence to item-writing best practices. Items were often reviewed simultaneously by an accessibility expert for adherence to principles of UDL, and for issues that students with disabilities or students who are English learners might encounter when accessing the item. The accessibility reviewer might refer items to experts with knowledge in low-incidence disabilities (e.g., blind or low vision, Deaf/Hard of Hearing) for further review. After these reviews, items that had undergone substantial changes were returned to the editing team. After the completion of internal reviews, items were sent to external committees and KSDE for review.

II.3.3. Grade-10 Mathematics Item Writing

The new grade-10 mathematics development occurred at an in-person item-writing event in June 2018 in Lawrence, and then with an item-writing vendor in fall 2020. In June 2018, the recruited item writers participated in item-writing training before the event. During the item-writing event, the items also underwent peer review. After reviewing items developed in the in-person item-writing event and the blueprint for grade-10 mathematics, AAI content experts identified gaps in

test blueprint coverage. True North Education Consultants was contracted to construct 35 assessment items to fill identified blueprint gaps in fall 2020. True North Education Consultants used three experienced item writers to write items for grade-10 mathematics. All item writers from True North had bachelor's degrees, and most had at least 15 years of teaching experience.

II.3.3.1. Item-Writing Event

II.3.3.1.1. Item Writers

Nine educators from across the state were invited to participate in the item-writing event in June 2018. The item writers were from Leavenworth, Manhattan, Olathe, Overland Park, Pittsburg, Shawnee, and Topeka, all in Kansas. Among these nine educators, 88.9% were female and 11.1% were male. Eight educators worked in public schools, and one educator worked in a private school. All of item writers had active teaching licenses and bachelor's degrees; two had master's degrees.

II.3.3.1.2. Item-Writer Training

Before writing items for the KAP assessment, item writers trained in the use of the KAP subject-area item specifications for writing and reviewing of items. All item writers received training in the following topics:

- accessibility, bias, and sensitivity considerations
  - accessibility
    - accommodations
    - differentiated assessments
    - language
    - students with disabilities
  - bias-and-sensitivity guidelines
    - advice
    - dangerous activities
    - inflammatory or controversial material
    - language inclusiveness
    - linguistic feature and language accessibility
    - population diversity
    - stereotypes
    - topic familiarity
  - equality versus equity
  - UDL
- alignment and resources
  - *Children's Writer's Word Book*
  - DOK framework
  - *EDL Core Vocabulary*
  - item specifications
  - mathematical practices
  - progression documents
  - standards and clusters in the 2017 Kansas Standards for Mathematics
- critiquing sample items

- differences in assessment types (e.g., formative, interim, and summative assessments)
    - student glossary
- item development
    - life cycle of an assessment
    - life cycle of an item
- item-writing best practices
    - answer-choice development
    - content, format, stem structure
    - traditional and nontraditional items

The following resources were shared with item writers to assist in writing items:

- DOK framework
- *EDL Core Vocabulary*
- *EDL Mathematics Vocabulary*
- *KSDE 2017 vs. 2010 Mathematics Standards Comparison Document*
- *KSDE Grade Level Focus Document*
- KSDE mathematics website
- KSDE progression documents for Mathematics Standards
- *KSDE Kansas Mathematics Standards Student Glossary*
- Secure item specifications
- *Wheel and Hess' Cognitive Rigor Matrix*

The item training also covered security of test materials, including the following requirements:

- Item writers must complete their nondisclosure agreement.
- If an item writer needs to leave the secure area, he or she must sign out a badge from an AAI staff member or be escorted by an AAI staff member.
- Item writers must leave workshop materials in room at all times. Secure testing materials must be shredded. At the end of the workshop, AAI staff will collect all materials.
- Test content and design discussions are confidential.

Item writers were provided multiple templates of item types available in the Kite® platform to construct their items, such as constructed response, matching lines, matrix, multiple-choice keyed, and multiple-choice–multiple-select. Writers were given guidance and options regarding graphic mockups as part of the process.

Item writers also practiced writing items during training. This practice included reviewing and evaluating previously written items and draft items. Item writers reviewed items that were rigorous and items that were problematic to identify and clarify the differences between items. After this review, the writers practiced correcting flawed mock items. Writers then began writing their own practice items for feedback from AAI staff and other participants. The open forum of the practice ensured writers were comfortable with the task.

II.3.3.1.3. Item-Writing Process

The item-writing event focused on clusters based on identified item-pool gap analysis. Item writers selected clusters of their greatest interest for item writing. Then, the pool gaps from the

pool analysis presented high-priority clusters for item writing that were lacking coverage. Finally, two to four educators were assigned to write to each of those clusters to fill coverage gaps.

All item writers followed a specific process as they developed items. First, an item writer wrote items to the metadata requirements, ensuring that the item followed the rules of item writing, the content was correct, any surrounding context was accurate, and the language was appropriate for the grade being tested; the item writer then verified the correct answers. Then, the item writer forwarded the completed item to a fellow item writer for review. The two writers discussed the item's alignment to the standards and its cognitive complexity demands. Next, the item writer reviewed and accepted necessary changes before submitting the item to the additional item review process.

Additional item-review process followed the submission of items. Item writers acting as peer reviewers were instructed to consider several questions:

- Does the item align to the appropriate cluster and standard?
- Does the item elicit evidence of student mastery for at least one standard in the cluster?
- What is the metadata for the item (variant ID, cluster, standard, DOK)?
- Is the key correct?
- Regarding bias and sensitivity,
  - Does the item create any barriers for student subpopulations?
  - Does certain student group have an advantage or disadvantage when answering the item?
  - Does the item include language or content that may be sensitive or offensive to a student group?

II.3.3.2. Additional Grade-10 Mathematics Item Writing

After the item-writing event, the AAI content team analyzed the available grade-10 mathematics items in the pool again after adding items developed from the item-writing event to the pool and discovered that additional items were still needed to meet the blueprint, which was finalized after the item-writing event. Thus, AAI contracted with an external writing vendor, True North, to develop the items needed to meet the blueprint requirements.

II.3.3.2.1. Resources and Process

AAI provided several resources to the external vendor, including style guides, graphic guidelines, and item templates.

The vendor's item-construction process included item writing, content review, editorial review, and a final review of item-writing guidelines. These four steps are described separately below.

For item writing, the vendor's content-area specialist wrote items tagged with contract-specific metadata (e.g., grade, content area, DOK, key or scoring guide, distractor analysis). The item writer also described required graphics and specified calculator usage or nonusage. For items with graphics, the item writer submitted graphic requests to the vendor's graphics team and reviewed the completed graphic for accuracy and adherence to specifications.

During content review, the lead item writer or a peer item writer reviewed the item, scoring guide, and graphics. The content reviewer considered item integrity, format, and content and structure; appropriateness to the designated content area; clarity, item key, and graphics quality. Fundamental questions for the content reviewer included, but were not limited to, the following:

- What does the item ask? Is it important? Does it align to the standard?
- Is the key the only possible key?
- Is the item complete (e.g., with content codes, key, grade, and contract identified)?
- For multiple-choice items, are the distractors viable and do they represent common errors and misconceptions?
- Is the item appropriate for the designated grade?
- As a set, do the items cover the blueprint?

After content review, the item writer and content reviewer resolved each element of the review, resulting in current versions of the items. The item then writer sent the revised set of items for the next editorial review.

For editorial review, the editor consulted and ensured compliance with the KAP Style Guide. The editor also considered item integrity, item format, item content and language, possible ambiguity, key, item bias, and issues related to sensitivity. In general, this editorial level of review involved, but was not limited to, the following actions:

- Eliminate confusing or vague wording, both from a conventional readability point of view and from the particular assessment point of view.
- Edit the wording of items.
- Ensure consistency and coverage of various responses in scoring guides.
- Suggest rewrites as necessary (final wording approved during item or item set resolution).
- Ensure consistency of usage and terms.
- Check for correct grammar and usage, and delete typographical and spelling errors.
- Consult *Developing and Validating Multiple-Choice Test Items* by Thomas M. Haladyna to ensure adherence to industry standards for multiple-choice items.

Then, the content-area specialist and editor resolved each element of the review.

For the final item review, the lead item writer or a peer item writer examined the final item and checked the following topics:

- General issues
  - The question clearly addresses the standard.
  - All content is accurate (graphics, passage, and question).
  - No economic, cultural, ethnic, gender, or religious bias is present.
  - Context is realistic.
  - Context and reading level are grade appropriate.
- UDL
  - Wording is clear and concise.
  - Syntax uses present tense and active voice when possible.

  - Reading level is as low as possible.
  - Simple sentence structure (subject–verb–object) is used.
  - Sentences are short.
  - No colloquialisms or words with double meanings are used.
- Item stems
  - Stem presents a definite, explicit, and singular question.
  - Stem is brief and free of irrelevant information.
  - Stem includes appropriate qualifiers (e.g., best, most likely) if necessary.
  - Stem is worded positively when possible.
- Item options
  - All choices are plausible.
  - Distractors capture common misconceptions or errors.
  - Numerical options are in ascending or descending order.
  - Answer choices are grammatically parallel (e.g., same part of speech, same sentence structure).
  - All choices are grammatically consistent with stem.
  - There are no clues to the correct answer (e.g., opposites, antonyms, synonyms, phrases repeated from stem).
  - All choices contain the same level of detail.
  - Answer choices are of about the same length (or two short choices and two long choices).

II.3.3.2.2. Internal Review and Revision Process

After AAI received the items, AAI mathematics content experts reviewed and revised them. The following guidelines were considered when reviewing the items:

- The stimulus and item are appropriate to the grade.
- KAP Style Guide specifications do not interfere with the content or functioning of the item.
- The item is free of content errors.
- The introductory text, stem, or prompt is appropriate and gives clear directions.
- The answer options provide a direct answer to the plain question asked in the stem.
  - Answer options are parallel to each other in language and substance.
  - Answer options are not needed to understand the meaning of the stem.
  - The test taker does not have to engage in process of elimination to determine the key.
- The key is accurate.
- Distractors are reasonable.
  - Distractors are plausible but incorrect.
  - No new information is presented in a distractor.
  - Rationales explain why the key is correct, and each distractor is plausible but incorrect.
  - Distractors do not use problematic wording.
- Items ensure equal opportunities for all students to demonstrate their knowledge, skills, and abilities.

Items that were revised after review were reviewed again using the guidelines described above to ensure changes to the items did not introduce any content or fairness concerns. AAI mathematics content experts who reviewed and revised these items included, but were not limited to, mathematics test-development coordinators with mathematics content expertise and accessibility-team specialists with special education and physical and sensory disabilities expertise.

II.3.4. Item Review

The item-review process involved several stages:

- internal content and editorial review
- external review (content and fairness) using multiple panelists
- internal content-team resolution
- data review
- accessibility review

The AAI content team performed the internal content review, after which the items went through editorial review. For items that needed graphics, a content-team member provided the graphic artist with instructions for rendering the stimulus and then confirmed that the completed graphic met the intended function. After the editor finished editing the items, the content team reviewed the items again before external review. If substantial changes were made to an item during the second round of internal content review, the content team returned the item to the editing team.

After completion of internal content and editorial reviews, the items went to external reviewers. The next section describes the reviewers and review process for two external reviews: content review and fairness review.

External reviewers reviewed ELA and mathematics grade 3–8 items developed in 2012 in 2013 in 2014. For ELA items developed in summer 2017, external review took place immediately after item writing. For grade-10 mathematics, external item reviews occurred in February 2021 for items developed from the AAI in-person item-writing event and vendor-produced items. External review of science items occurred twice as asynchronous online events: from spring to winter in 2015, and from fall to winter in 2016.

AAI staff then considered the items recommended by educators during the external review, incorporated edits to the items, and presented the items to KSDE for final approval. AAI staff consulted with KSDE about whether the internal editing as needed. Next, items were field tested, and student-response data were used in various analyses. AAI psychometricians and the content lead reviewed the field-test data analysis during data review.

For items passing the data review and placed on the operational assessment or used as operational field-testing items, accessibility reviews were conducted for adherence to principles of UDL and for issues that students with disabilities or English learners may encounter when accessing the item. Data and accessibility reviews are discussed in Section II.3.4.3 and Section II.3.4.4, respectively.

II.3.4.1. External Item Reviewers

AAI and KSDE staff recruited Kansas educators to serve as item reviewers for two separate types of reviews: content review and fairness review. Prospective external item reviewers for ELA, mathematics, and science completed an online survey in which they indicated their demographic information, teaching experience, professional qualifications, content expertise, experience with the standards, and special education or English learner.

Content review panels for ELA and mathematics grades 3–8 were formed by grade band: grades 3–5, grades 6–8, and high school. Of 36 reviewers who participated in the science item external review in, 22% were male and 78% were female 2015 and 29 reviewers with 21% male and 79% female participated the science external review in 2016. Content-review panels for science are formed by grade, but some reviewers served in more than one panel because domain content knowledge often extends above or below grade. Bias-and-sensitivity panels were assembled and included members of various groups to reflect the diversity of Kansas and represent a number of minority groups. Item reviews were processed through a secure, online reviewing system. After completing a web-based training session, reviewers evaluated items at their own pace and provided feedback by a given deadline.

For the grade-10 mathematics external item review, educators who were involved in the item-writing workshops noted above were not eligible to participate in the external review. AAI staff provided the list of potential reviewers to KSDE staff for selection to each panel. Nine panelists were chosen to participate in the content review, and seven panelists were chosen to participate in the fairness review. The demographic information for the content review panel and for the fairness-review panel is summarized in Table II-7 and Table II-8. The content review panel consisted of half classroom teachers and half nonclassroom teachers from different State Board districts. Most content-review panelists had more than 10 years of experience. Most fairness-review panelists were classroom teachers from different State Board districts. About half of the educators had fewer than 10 years of experience, and the other half had more than 10 years of experience.

Table II-7. Demographic Information of Grade-10 Mathematics Content Review Panelists

Characteristic	%
Sex	
Female	78
Male	22
Race	
White	78
Black	0
Asian	11
Native American	11
Other	
Ethnicity	
Hispanic	0
Non-Hispanic	100
Role	
Classroom Teacher	56
District Staff	33
Other	11
SBOE District	
1	11
2	11
3	0
4	11
5	22
6	11
7	11
8	0
9	22
10	0
Years of experience	
< 3	0
3–5	0
6–9	11
≥ 10	89

Note: SBOE = School Board of Education.

Table II-8. Demographic Information of Grade-10 Mathematics Fairness-Review Panelists

Characteristic	%
Sex	
Female	86
Male	14
Race	
White	100
Black	0
Asian	0
Native American	0
Other	0
Ethnicity	
Hispanic	0
Non-Hispanic	100
Roles	
Classroom teacher	86
District staff	14
Other	0
SBOE district	
1	14
2	0
3	29
4	0
5	14
6	14
7	14
8	0
9	14
10	0
Years of experience	
< 3	0
3–5	29
6–9	29
≥ 10	43

Note: SBOE = School Board of Education.

II.3.4.2. External Item Review

The external reviews included an orientation and asynchronous review of the items. The grade-10 mathematics external review also included a synchronous discussion of the items after the asynchronous review. The next sections include details about each stage of the external review. Orientation and the synchronous discussions occurred via an online meeting platform, and the asynchronous review occurred through a secure, online reviewing system. After completing the orientation, reviewers evaluated items at their own pace and provided feedback by a given

deadline. Then, the panels discussed all feedback and possible revisions for items during the synchronous discussion for grade-10 mathematics only.

II.3.4.2.1. External Item-Review Orientation

All item reviewers participated in an orientation for either the content-review or fairness-review. Orientations included two components: one session of specialized training for content or fairness review and another session for the online review system. Both the content and fairness orientations included security reminders, background information, and an overview of the assessment-development process. After the orientation, panelists were encouraged to ask questions about their review responsibilities. Panelists were trained in the Review Management System and practiced item review to familiarize themselves with the review platform. After panelists confirmed their confidence with the review and system, they engaged in the asynchronous review. Panelists were given contact information of AAI staff in case of additional questions.

II.3.4.2.2. Item Content-Review Process

For the asynchronous review of content, panelists independently reviewed items and were given dates by which to submit their ratings and comments. Content reviewers considered every aspect of each item: alignment to content standards, appropriateness (i.e., content, context, and vocabulary for the grade and subject), correct and incorrect answers, and utility and clarity of graphics or stimulus.

In general, content reviewers checked items for

- alignment to clusters or targets
- grade appropriateness, including language and context
- content errors

After analyzing items, reviewers recommended that they be accepted, revised, or rejected. For items that were revised or rejected, panelists gave specific reasons (e.g., “item aligns better to this cluster”).

Next for grade-10 mathematics, AAI staff saved all panelist ratings and comments in preparation for the content synchronous discussion; 43 of the 125 revised or rejected items were included in the content synchronous discussion. During the discussion, panelists explained their concerns about items and suggested solutions, such as clarifying language, changing the format, and editing options to prevent multiple keys. AAI content experts considered these suggestions for item revisions.

II.3.4.2.3. Item Fairness-Review Process

For the asynchronous review of fairness, panelists independently reviewed items and were given dates by which to submit their ratings and comments. Fairness reviewers identified barriers not related to content standards that could prevent students from demonstrating what they know and can do. These barriers include unfamiliar language; linguistic complexity; potentially sensitive topics; stereotypes (both positive and negative), including emotions, regions, and occupations; accessibility for special populations; and issues with cultural or experiential knowledge.

In general, fairness reviewers checked items to

- identify potential bias and sensitivity
- ensure all items were appropriate and accessible for all Kansas students, including
  - principles of UDL incorporated into items
  - appropriate language complexity for all students
- ensure representation that broadly and generally reflects the student population

After analyzing items, reviewers recommended that they be accepted, revised, or rejected. For items that were revised or rejected, panelists gave specific reasons.

Next for grade-10 mathematics, AAI staff saved all panelist ratings and comments in preparation for the fairness synchronous discussion; 102 of the 125 revised or rejected items were included in the fairness synchronous discussion. During the discussion, panelists explained their concerns about items and suggested solutions, such as removing extraneous language, adding clarifying language, adding graphics, and simplifying directions. AAI content experts considered these suggestions for item revisions.

After the content and fairness external review, AAI content experts revised items and presented them to KSDE staff for approval for field testing.

II.3.4.3. *Data Review*

After field-test or operational-field-test item analysis and before test construction, psychometricians and content leads reviewed item statistics. Items with statistical flags were only used when the item pool did not have other items for blueprint coverage. Item statistical flagging criteria are explained in Appendix B. When flagged items were used as operational items, they underwent extra review and discussions.

II.3.4.4. *Accessibility Review*

After content leads and psychometricians identified items for form construction according to blueprint and psychometric specifications, an accessibility expert added accessibility features to ensure the widest range of students can access the items. The accessibility enhancement incorporated knowledge of disabilities (e.g., blind or low vision, Deaf/Hard of Hearing, English learner status). Every item that had not previously appeared on an accessible version of a form underwent review before placement on an operational form.

Accessibility features that were incorporated into items include

- accessible color palettes
- appropriate color contrast settings
- alternative text on images
- keyboard navigation
- compatibility with commonly used assistive technology products, such as screen readers
- braille
- key word translations
- American Sign Language (ASL) videos
- text-to-speech

II.3.5. *Field Testing*

For all three subjects, field-test items were embedded in the operational test forms and were field tested for future KAP assessments. All subjects and all grades have field-test items except grade-11 science. Table II-9 displays the number of field-test items by subject and grade.

Table II-9. Number of Field-Test Items by Subject and Grade

Grade	English language arts	Mathematics	Science
3	41	57	—
4	40	55	—
5	49	59	32
6	39	51	—
7	41	63	—
8	46	53	36
10	52	55	—
11	—	—	0

II.3.6. Field-Test Data Analysis

Field-test item analyses included classical item analysis and differential item functioning analysis. Items that were too easy or too difficult, that did not discriminate students' ability well, or that had large differential item functioning were flagged according to predetermined criteria (Appendix B). Flagging statistics will be used in future data review and test construction.

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 *Kansas Assessment Examiner's Manual 2021–2022* (*Examiner's Manual* hereafter). The *Examiner's Manual* provides information regarding standardized test administration for districts, schools, and teachers. It also provides guidance on the administration procedure for the 2021–2022 KAP assessment. Main topics of the *Examiner's Manual* include

- overview of KAP assessment
- test security and ethics
- accommodations
- preparation activities before test administration
- directions for test administration on testing day
- activities for after test administration
- resources for test administration

For all subjects, grades, and students, KAP is entirely computer based, and the delivery platform is Kite Student Portal (described in Section II.4.2. Test-Administration Procedures). To take KAP assessments, Student Portal must be installed on students' computing devices. The 2022 KAP testing window opened on Monday, March 21, 2022, and closed on Friday, April 29, 2022. Each test session was designed to take approximately one class period (i.e., 45–60 minutes).

Thus, one test was designed to take approximately two class periods. However, all KAP tests are untimed, as enough time should be given to students to finish testing.

II.4.1. Test-Administration and Security Training

Kansas uses a train-the-trainer model, in which District Test Coordinators (DTCs) receive training directly from KSDE and then train building-level personnel before local test administration. First, the test-administration and security trainings for all Kansas DTCs include: test-security and ethics training, DTC virtual training webinars, and DTC or building test coordinator (BTC) regional training held in different locations in September and October. Then, DTCs train local test administrators.

For DTC training, test-security and ethics training is offered as online training modules by KSDE. All DTCs must participate yearly in one module. After training, DTCs must verify training and agree to adhere to policies and practices in the training. For 2021–2022 test administration, all DTCs needed to complete and verify the training module before November 30, 2021. The test-security and ethics training covers test-security procedures, test-administration monitoring, roles and responsibilities, reporting testing discrepancies and potential violations, reporting item issues, the security of personal identifiable information (PII), accommodations, and appropriate and inappropriate testing practices. Details about test-security and ethics training can be found in the Kansas Test Security and Ethics training slides. DTC virtual training webinars were held on the second Tuesday of each month in 2021–2022. DTCs who could not attend live webinars could access online training materials and webinar recordings on the KAP DTC Virtual Training Webinars website at any time. The trainings were offered by AAI in partnership with KSDE. The trainings provided updates on KAP and Kite technology, an overview of important training dates, a description of accommodations, directions for ordering braille booklets, and updates on special circumstance codes. The regional trainings were in-person trainings and covered test coordinator responsibilities, using Kite, and updates for the upcoming year. The regional trainings also offered test-security and ethics training as another opportunity for DTCs to participate.

For local training, DTCs train staff members who administer state assessments at the district or building level before testing begins. Local staff members include administrators, educators, paraeducators, and other appointed staff members. The training includes test security and ethics, reporting, and accommodations. For all training at the district and building level, DTCs document the personnel, time, and method of the training, and maintain records at the district and building levels. Anyone administering a KAP assessment had to complete all district- and building-level training by March 18, 2022. After completing training, staff administering state assessments signed an agreement to abide by state ethical testing practices and provide written verification.

II.4.2. Test-Administration Procedures

The *Examiner's Manual* includes guidelines for administering KAP assessments in a standardized and secure procedure; KSDE developed and approved these guidelines. All test administrators are required to read the *Examiner's Manual*. The standardized and secure test-administration procedures before, during, and after KAP administration are described in the next sections, and more-detailed information can be found in the *Examiner's Manual*. Further details

about administration-related accommodations can be found in Chapter V. Inclusion of All Students of this manual.

II.4.2.1. Before KAP Administration

Before KAP administration, local testing windows should be scheduled to ensure all students can finish testing before the end of the school day and before the end of the testing window. Once the local testing windows are scheduled, those dates should be added in Kite Educator Portal. Districts can then register students for testing and submit students' records so they can use Kite Student Portal. Also, teachers should complete the Personal Needs and Preferences Profile settings for students who need accommodations and enter special circumstance codes for students who cannot take KAP assessments.

As the local testing window nears, test administrators should

- prepare the room for testing (e.g., remove instructional material that may give clues)
- have appropriate manipulatives for the mathematics and science assessments
- be familiar with rules for using resource sheets and calculators for mathematics assessments
- have students' individual usernames and passwords ready
- have access to Daily Access Codes (DAC; needed to access KAP assessments)
- have needed materials ready (e.g., pencils, scratch paper, clocks, and headphones)

To better prepare students for KAP assessments, educators should strongly encourage students to take the practice tests. The *KAP Practice Test Guide for Educators* is available for educators to support students' access to practice tests. Kite Student Portal provides practice tests to help students gain confidence navigating assessments and become familiar with different KAP item types before taking the test. There are two types of practice tests: a technology practice test that includes various item types using simple content, and a subject-oriented practice test that features various item types with subject-oriented content. The subject-oriented practice test also provides a deeper look at different tools. All practice tests are grade banded: the technology practice test includes grade bands K–1, 2–5, and 6–12; and the subject-oriented practice test includes grade bands 3–5, 6–8, and 10–11. Technology practice tests are not secure and should be used to help students gain experience taking assessments on the online platform (i.e., Kite Student Portal) and feel confident taking the actual KAP assessments.

II.4.2.2. During KAP Administration

On assessment day, test administrators make sure students are taking the correct test, help students log in to Kite Student Portal, instruct students to enter the DAC, and remind students not to disrupt others if they finish early. After testing starts, test administrators should only read specific scripts provided in the *Examiner's Manual* as instructions. To ensure a quiet testing environment and help students focus on testing, test administrators also need to follow proctoring guidelines given in the *Examiner's Manual* during testing. Last, before a student exits the test, test administrators should verify the review screen to ensure all items were answered.

II.4.2.3. After KAP Administration

After one KAP test session administration, test administrators should collect all materials, such as manipulatives for mathematics and science. Also, test administrators should collect and

destroy all materials, including scratch paper. Then, DTCs monitor student testing status and reactivate student testing sessions if needed.

II.5. Monitoring Test Administration

Test-administration monitoring includes monitoring both testing and testing data. Testing monitoring also includes both local monitoring and KSDE visits. For local monitoring, DTCs can monitor students' test progress, such as test-session status, via Kite Educator Portal. Building principals, BTCs, and DTCs can also monitor item status in each test session in real time using Kite Educator Portal. DTCs and other test administrators are responsible for identifying any testing discrepancies; testing discrepancies are any violations of standard test-administration procedures. After testing discrepancies are identified, superintendents or their designees are responsible for reporting them in writing.

Every year during the testing window, KSDE staff and members of the Kansas Assessment Advisory Council visit approximately 5%–10% of Kansas schools to monitor administration and test security. The schools are selected through either volunteering or random selection. However, because of COVID-19, on-site visits from KSDE and members of the Kansas Assessment Advisory Council were halted in 2022.

During the operational window, monitoring of testing data was conducted by Agile Technology Solutions (ATS), a center of AAI that oversees and manages the Kite system, and the AAI psychometric team. ATS conducted data validation daily to monitor system usage and identify testing irregularities. System usage includes a DTC training log, click history of student responses, test-taking hours, test-status summary, server load, the number of Kite Service Desk (i.e., support for Educator Portal and Student Portal) tickets, and the frequency of test reactivations. Testing irregularities include fast test-taking behavior (i.e., students finished a test section in a short amount of time), irregular testing time (i.e., a test session started or ended outside of school hours), tests reactivated by users (i.e., test administrators) or by the system, and student enrollment or demographic data error.

The AAI psychometric team periodically conducted student-response data checks to ensure quality administration. Those checks included verifying that

- student demographic information was entered in Kite
- student test information values were accurate
- students received only one score for each item
- item scores matched the possible item scores
- all possible item scores were obtained by at least 1% of students
- each student had only one set of demographic information
- each student took only one test form in each subject
- distributions of demographic information and test information were reasonable
- students' raw scores in the first session had a strong relationship with those in the second session
- frequency distribution of students' raw score was smooth, bell shaped, and generally increased then decreased as raw scores increased

II.6. Test Security

Test security focuses on several important facets: test materials, test-related data, PII, and accommodation-related security. Test security should be protected through the whole testing cycle, from test development and administration to scoring and reporting. Moreover, to protect the security of all facets, both physical security and online, platform-security requirements should be met and strict procedures should be in place during administration and reporting.

The electronic item bank, online administration system, and student responses are stored in the Kite platform, which is designed and maintained by ATS. Three portals were designed within the Kite platform to serve different needs:

- Content Builder, for item and test development
- Educator Portal, for educators to input and access test and student information
- Student Portal, for online testing

The Kite platform uses Amazon Web Services (AWS) in high-availability mode with no single point of failure. Using AWS ensures that loss of any given server loss or even of an entire availability zone (i.e., data center) will have minimal impact on Kite platform availability. Recovery times are very short, ranging from no downtime (for loss of most servers) to a few minutes (for loss of an entire data center). Moreover, AWS fully managed the recovery, which runs in high-availability mode and is automatic. Using a service provided by AWS, the Kite platform has a multilayered design to prevent denial-of-service attacks and system intrusion. The Kite platform moved to AWS in 2017. Since then, the Kite platform has experienced no outages that have affected testing.

KSDE has predetermined procedures to deal with testing discrepancies and possible security violations. All testing discrepancies and possible security violations should be reported to KSDE. Upon breach of security, appropriate consequences are put in place at the district level. Depending on the uniqueness of each case, possible steps vary and may include, but are not limited to,

- no action because the breach was not severe enough to warrant any action
- KSDE action, such as a written letter or phone call to the superintendent or DTC, stating concerns and monitoring action steps
- retesting of students
- removal of test proctors from testing rooms
- KSDE follow-up monitor visits the next testing year to ensure changes to inappropriate practices have been made

For more details, refer to the *Kansas Assessment Fact Sheet: Test Security and Ethics*.

II.6.1. Test-Materials Security

To protect test-materials security during test development, the physical security requirements are met by using hosting providers that conform to the Statement on Auditing Standards (SAS-70) for physical access and PCI Data Security Standard compliance. Most activities related to project management, test development, and data analysis take place at Accessible Teaching, Learning, and Assessment Systems (ATLAS; also a center of AAI) and ATS. Both centers are in secure wings that can be accessed only with a key or key card. In general, work is done either at one of

our centers using secure server systems or a secure virtual private network connection. Moreover, the electronic item bank stored in Kite Content Builder can be accessed only by authorized users of Kite. For any activities involving external item reviewers, such as Kansas educators, all participants are required to sign nondisclosure agreements to ensure item and task confidentiality and security.

To protect the test-materials security during test administration, specialized training and certification for test administrators are required (described in Section II.4.1. Test-Administration and Security Training). 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 in Educator Portal each year. By accepting the security agreement, users agree to not store or save assessment materials to computers or personal storage devices, to not print assessment materials, and to not share personal passwords with others.

II.6.2. Test-Related Data Security

For test administration, all Kite portals 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 portals generate access records that system administrators can review to track access. Access to individual Kite portals is controlled according to established policies for that application and the data it maintains. All access policies and accounts are reviewed periodically to ensure that access to systems is limited to the appropriate populations.

DTCs attend the test-security and ethics training provided by KSDE and oversee test security for the entire district. They establish procedures that determine which appropriate personnel can access Educator Portal and their role assignments within the district. DTCs also remove or deactivate from Educator Portal any users who leave the district or change roles within a district. Moreover, DTCs establish and describe processes ensuring the usernames and passwords in Educator Portal are exclusive to the users and confirm that users' rights are permitted according to their roles.

II.6.3. Security of Personally Identifiable Information

In accordance with the Family Educational Rights and Privacy Act (FERPA), students, teachers, operators, and administrators who have access to personal student data are limited to only the student records in which they have a legitimate educational interest; all users are provided minimal necessary access. Throughout each school year, security levels, groups, and access are reviewed periodically to ensure continued compliance.

All test administrators are informed that PII should not be conveyed when reporting testing issues. The documentation for Kansas regarding allowable identifiers in an email specifies that only the State Student Identifier, and no other identifying details (e.g., name, district, school) should be provided in an email. In cases when the Kite Service Desk needs to be contacted, students' PII cannot be sent via email or live chat.

For scoring and reporting purposes, students' PII data are stored on secure servers in AAI. AAI staff working on KAP materials and who have access to PII data on servers are required to complete annual KSDE information-technology security and data-privacy training to ensure

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

After scoring and reporting are complete, ATS provides student-assessment data (e.g., return files, score reports) to KSDE. Those data are placed on a secure drive that only specific members of the ATS and KSDE teams can access. For school and district reporting, scores from more than 10 students are needed for aggregated results; this is to prevent identifying individual students' scores. Descriptions of KAP results in technical documentation are reported only at the aggregated level.

II.6.4. Accommodations-Related Security

Local staff members who administer a state assessment must complete the test-administration and security training given by DTCs, sign an agreement to abide by state ethical testing practices, and provide written verification of training before local testing begins. The training covers the ethics of testing, test security, and reporting and documenting accommodations. To ensure security related to accommodations, DTCs need to establish procedures for entering student accommodations in the Personal Needs and Preferences Profile in Educator Portal and keep documentation for text-to-speech accommodations and other accommodations that require deviating from general administration of the assessment. More information about selecting and entering information in the Personal Needs and Preferences Profile is in Section V.4.1. Selection of Accommodations. Text-to-speech accommodation of ELA passages must be approved by KSDE before testing. Thus, either DTCs or BTCs need to submit the need for text-to-speech accommodation of ELA passages to KSDE at the beginning of the year. During the assessment, Kite audio (i.e., headsets) is used for text-to-speech accommodation rather than a human reader.

III. Technical Quality: Validity

According to 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 [APA] et al., 2014, p. 11).

The *Standards* (APA 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 test content, response processes, internal test structure, relationships among test scores and other variables, and the consequences of testing. The validation process involves the ongoing collection of a variety of evidence to support the proposed test-score interpretations and uses. This chapter mainly describes aspects of the Kansas Assessment Program (KAP) assessments that support KAP test-score interpretations and uses.

Because validity evidence supports the intended uses of test scores, it is necessary to identify the purposes of a test before providing evidence to support test validity. The purposes of the KAP assessment, as described at the beginning of this manual, include (a) measuring specific claims related to the Kansas Standards, (b) reporting students' academic performance, and (c) using local assessment scores to assist in improving educational programs in the three subject areas (i.e., English language arts [ELA], mathematics, and science).

The gathered evidence on test content, response process, and internal structure supports the use of the KAP assessment to measure the Kansas Standards 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 reporting students' academic performance. Validity evidence from other sources, such as comparing KAP results with National Assessment of Educational Progress (NAEP) results, uses additional data to validate the use of KAP test scores.

III.1. Validity Evidence Based on Test Content

Validity evidence based on test content refers to how well test content related to specific content domains match what the test was intended to measure. Content evidence for KAP assessments comes from the alignment between KAP items and the Kansas Standards, from the congruence between the test and the test blueprint, and from the congruence between the test blueprint and the Kansas Standards (i.e., a balance of representation of standards). Content specialists at the Achievement and Assessment Institute (AAI) followed several steps to evaluate the content validity of the KAP assessment.

- Develop the test blueprint and specification and evaluate the relationship between the blueprint and the Kansas Standards.
- 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.
- Evaluate the alignment between KAP assessments and the Kansas Standards.

- Evaluate the degree to which the assessment addresses the depth and breadth expectations of the Kansas Standards in terms of the blueprint.

Chapter II Assessment System Operations presented validity evidence related to the development of the test blueprint, item and test development, and item review. As described in those chapters, the KAP blueprint has the same structure as the Kansas Standards. Test content specialists developed and aligned all KAP items with the Kansas Standards, and item development followed well-established procedures. After item development, items underwent multiple rounds of content and bias reviews. After field-test administration, psychometricians and content specialists reviewed the items' statistical properties, evaluating items from content and psychometric perspectives before selecting items for operational use. Districts then administered KAP assessments according to standardized procedures and provided accommodations for students with special needs.

The following list summarizes the efforts to ensure content validity.

- The development of the blueprint is a collaborative process between AAI, the Kansas State Department of Education (KSDE), and educators in Kansas. The blueprint uses the same framework as the Kansas Standards, ensuring the range and variety of standards measured in KAP are appropriate, as indicated in Section II.2.1. Test Blueprints.
- The proportion of items for each classification or claim in Table II-5, i.e. test blueprints, show that each classification or claim has an adequate number of items to represent the knowledge and skills described in the Kansas Standards.
- AAI and KSDE selected and trained qualified item writers to ensure they write high-quality items.
- AAI and KSDE established detailed item-development guidelines to train item writers, who also participate in guided item writing.
- AAI content specialists and editors review each new item and consider grade appropriateness, graphics, grammar and punctuation, language demand, and distractor reasonableness.
- External content reviewers review each item to make sure all items align with the Kansas Standards. They also consider grade appropriateness; verify correct answers; evaluate incorrect answers; and assess the need, utility, and clarity of any included graphics or stimulus.
- External bias, fairness, and sensitivity reviewers review items to identify barriers that may prevent students from demonstrating what they know and can do when those barriers are not related to the content standards.
- Before items are selected for operational use, both AAI psychometricians and content leads review the results of items' classical item analysis and distractor analysis to prevent items with extreme statistics from being used on operational forms.
- AAI accessibility experts review each new item to make sure the widest range of students can access the items.
- Standardized administration of KAP assessments minimizes the effect of the variation of administration and provides accommodations for students who need them. Students are given ample time to complete the tests to avoid speediness issues.

The validity evidence related to alignment between KAP items and the Kansas Standards, as well as the degree to which the assessment addresses the depth and breadth of the Kansas Standards in terms of the blueprint, come from an alignment study conducted by an independent external vendor. Several alignment studies occurred at different times to collect validity evidence related to alignment for 2022 KAP assessments. The next sections summarize the study procedures and findings from different alignment studies. All studies indicate strong or moderate alignment between KAP assessments and Kansas Standards.

III.1.1. English Language Arts and Mathematics Grades 3–8 Alignment

From fall 2014 through spring 2016, edCount conducted several rounds of reviews to evaluate two kinds of alignment: alignment between the KAP ELA and mathematics item pool and the Kansas Standards, and alignment between the blueprint and the Kansas Standards (Forte et al., 2016). Different from typical alignment studies that are designed for post-hoc evaluation, edCount used Forte's (2013, 2016) framework to develop a process that includes items from past administrations in the early evaluation stage and emphasizes the alignment among item, blueprint, and content standards.

For item-pool alignment, six ELA panelists and six mathematics panelists reviewed approximately 355 ELA items and 234 mathematics items of the 2016 KAP operational test. The panelists evaluated whether each item clearly and accurately reflects the content target (i.e., a group of standards) and depth of knowledge (DOK) levels recorded by item developers. Section IV.3.3. Cognitive Complexity describes different levels and ranges of DOK. The panelists' ratings indicated more than 78% of ELA items and more than 56% of mathematics items were consistent with intended targets across grades. For DOK, more than 54% of ELA items and more than 48% of mathematics items were rated by the panelists with a DOK level consistent with intended DOK level across grades. When DOK levels did not match, panelists' DOK ratings were usually higher than the intended DOK levels

The edCount blueprint-review panel, composed of four internal content and research staff members, used the internally developed protocols to assess the connections among the Kansas Standards, the KAP test blueprint, and the item pool. 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 (i.e., a group of targets) 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 the blueprint expectations. Because KAP did not have DOK blueprints, averages of DOK by target were computed. ELA DOK was 2.4 for all grades, and mathematics DOK ranged from 2.4 to 2.6. The values indicated more items in the level-3 DOK (i.e., higher cognitive complexity).

In summary, the edCount concluded the alignment of the KAP assessment with Kansas Standards was strong across item pools and blueprints. This conclusion indicates that item and test development resulted in assessments that strongly reflect the content expectations laid out in the content documents, with some exceptions. To address those exceptions, edCount provided some recommendations.

- Some items have off-grade alignment to content standards, and the intended targets should be adjusted.

- Blueprints should include DOK requirements.

AAI adjusted the KAP assessments according to edCount's recommendation.

- The realignment in 2017 modified the grades of some items.
- The blueprint published in the online resource document in 2018 have included the DOK requirements.

III.1.2. Grade-10 Mathematics Alignment

EdMetric conducted an independent external study with Kansas educators in July 2022 to examine the extent of alignment between the KAP grade-10 mathematics assessment and the 2017 Kansas Standards for Mathematics (Egan & Davidson, 2022a). The purpose of the study was to examine

- the extent of alignment between the KAP grade-10 mathematics assessment and the 2017 Kansas mathematics high school standards in terms of content (i.e., knowledge and process), balance of content, and cognitive complexity
- the extent to which the KAP grade-10 mathematics assessment addresses the depth and breadth of the 2017 Kansas mathematics high school standards in terms of the blueprint

Thus, Kansas educators evaluated the alignment between the KAP grade-10 mathematics assessment and the 2017 Kansas Standards for Mathematics (i.e., content standards), and the alignment between the assessment and the blueprint. As mentioned in Section II.1. Assessment Framework of the Assessed Grades, the Kansas Standards for Mathematics organize the standards for grade-10 mathematics into domains, conceptual categories, and classifications. Alignment is evaluated at the conceptual category level.

Eight educators, with an average of 13.5 years of experience in teaching, participated in a one-day virtual workshop on July 19, 2022, to evaluate alignment of 56 grade-10 mathematics items and the complete operational form used for scoring and reporting for the 2022 administration. The educators represented different regions of Kansas and were diverse in gender, race, and urban or rural location. Before the workshop, a content expert from EdMetric independently assigned alignment ratings to assessment items and standards.

The workshop started with alignment training for panelists, after which panelists indicated they understood the process and their roles before they started rating. Two groups of panelists reviewed the alignments made by EdMetric's content expert and independently made changes. Within the groups, panelists discussed any disagreement on alignment ratings after individual ratings. Next, the two groups came together and discussed any disagreement. For the whole alignment study, panelists indicated that they could provide rationales for their ratings (Egan & Davidson, 2022a).

EdMetric used the modified Webb approach (Webb, 1997, 1999) to evaluate the alignment of items to content standards and the blueprint; this method is efficient and can be implemented easily. The evaluation included four criteria: categorical concurrence, DOK, range of knowledge (ROK), and balance of representation (BOK). The next section summarizes the alignment results for both the content standards and blueprint using these four criteria.

III.1.2.1. Categorical Concurrence for Grade-10 Mathematics

Categorical concurrence refers to the degree of similarity and consistency between the standards and assessment content. The categorical-concurrence criterion evaluates (a) the number of items per conceptual category (for alignment between assessment and content standards), and (b) the differences in item distribution between the assessment and the blueprint (for alignment between assessment and blueprint).

For each grade-10 mathematics conceptual category, results indicated that both content standards and blueprints have strong alignment with items according to the categorical-concurrence evaluation criterion. This result means at least six items in grade-10 mathematics aligned to each conceptual category. Differences between the expected blueprint percentage and the actual percentage based on panel ratings were smaller than 5% for all conceptual categories.

III.1.2.2. Depth of Knowledge for Grade-10 Mathematics

Depth of knowledge (DOK) evaluates alignment of cognitive complexity between assessments and standards. The content expert from EdMetric rated the target range of DOK for each standard in the 2017 Kansas Standards for high school mathematics. The blueprint also specifies the DOK goals as Level 1–2 for 75%–88% items of the skills-and-concepts classification and as Level 2–3 for 12%–25% items of strategic-thinking-and-reasoning classification.

Panelists rated the DOK levels of each item. EdMetric compared these item-level DOK ratings with the content-standard DOK target ranges and blueprint DOK goals. The DOK criterion evaluates (a) the percentage of items per conceptual category at or above the target DOK ranges (for alignment between assessment and content standards), and (b) the differences in DOK distribution between the assessment and the blueprint (for alignment between assessment and blueprint).

The results indicated that items were strongly or moderately aligned by conceptual category to the content-standard DOK target ranges. However, only the numbers-and-quantity-and-algebra conceptual category had strong alignment between DOK goals specified by the blueprint and item-level DOK. For all other conceptual categories, most items aligned to DOK Level 2 or 3 instead of the goal DOK Level 1 or 2.

III.1.2.3. Range of Knowledge for Grade-10 Mathematics

Range of knowledge (ROK) evaluates the extent to which the assessment covers the standards (Webb, 1997). The evaluation criterion examines the percentage of domains measured by at least one item in each conceptual category for both content standards and blueprint.

For each grade-10 mathematics conceptual category, results indicated that both content standards and blueprint had strong alignment with the assessment according to the ROK evaluation criteria; more than 60% of domains were measured by at least one item in each conceptual category for both content standards and blueprint.

III.1.2.4. Balance of Representation for Grade-10 Mathematics

Balance of representation (BOR) examines how items are distributed across the standards. This alignment criterion examines whether items in a conceptual category are evenly distributed across the domains within that conceptual category for both content standards and blueprint.

Results indicated that the geometry conceptual category was not aligned, functions and numbers-and-quality-and-algebra conceptual categories were moderately aligned, and other conceptual categories were strongly aligned between the assessment and the content standards for BOR. For BOR alignment between blueprint and assessment, all conceptual categories were strongly or moderately aligned.

In summary, the results from the external-alignment study suggested that alignment between assessment and content standards were strong or moderate for categorical concurrence, DOK, ROK, and BOR for all conceptual categories except one (i.e., geometry). The geometry conceptual category showed no BOR alignment between assessment and content standards, but strong BOR alignment between assessment and blueprint. Moreover, assessment and blueprint are strongly or moderately aligned for categorical concurrence, ROK, and BOR for all conceptual categories but are not aligned for DOK for most conceptual categories. On the other hand, all conceptual categories are strongly or moderately aligned for DOK between assessment and content standards.

III.1.2.5. AAI Response to Grade-10 Mathematics Alignment Study

There are different results on geometry BOR criterion for the content standard alignment and for the blueprint alignment. These differences are because there are fewer domains included in the test blueprint than the content standards. When educators constructed the grade-10 mathematics blueprint, they prioritized the domains that were more appropriate for the state standardized assessment and deemphasized other domains that were more appropriate for formative assessment. For example, educators thought the geometry domain "modeling with geometry" (G.MG) would be appropriate for the formative assessment. Thus, some domains were not included in the blueprint, like the geometry domain "circles" (G.C) and "modeling with geometry" (G.MG); these differences explain why the geometry conceptual category has no BOR alignment from the content-standard perspective but strong BOR alignment for the blueprint.

The lack of alignment on DOK between assessment and blueprint is caused by the process of determining the blueprint DOK requirement. When the blueprint was developed, educators considered the DOK range at the classification level rather than at the individual standard level. Aggregating the DOK range at the individual standard level would lead to a higher range on the skills-and-concepts classification than the current requirement. It would be beneficial to have educators evaluate the DOKs of individual standards, then aggregate standard DOK to the skills-and-concepts classification, which would lead to the blueprint DOK requirement falling into the range reflecting the standard DOK targets.

III.1.3. Science Alignment

EdMetric also conducted an independent external study with Kansas educators in September 2022 to examine the extent of alignment between KAP science assessments and the 2013 Kansas Standards for Science (Egan & Davidson, 2022b). The purpose of the study was to examine

- the extent of alignment between the KAP science assessment and the 2013 Kansas Standards for Science in terms of content (i.e., knowledge and process), balance of content, and cognitive complexity

- the extent to which the KAP science assessment addresses the depth and breadth of the 2013 Kansas Standards for Science in terms of the blueprint
- the extent to which the KAP science assessment address the three dimensions of science (i.e., Disciplinary Core Ideas [DCI], Scientific and Engineering Principles [SEP], Crosscutting Concepts [CCC]) as defined by the 2013 Kansas Standards for Science.

Thus, Kansas educators evaluated alignment between the KAP science assessments and the 2013 Kansas Standards for Science (i.e., content standards), alignment between assessment and blueprint, and multidimensionality measured by items on the assessments. As mentioned in Section II.1. Assessment Framework of the Assessed Grades, the Kansas Standards for Science organizes the standards into targets and domains and the KAP science assessment organizes the standards into targets and claims. The domains of the standards are Earth and space science, engineering, life science, and physical science. The claims of the blueprints are Earth and space science, life science, and physical science. The alignment is evaluated at the domain level for the content standards and at the claim level for the blueprint.

Fifteen educators, with an average of 18.5 years of experience in teaching, participated in a two half-day virtual workshop September 13–14, 2022, to evaluate alignment of 115 science items and the complete operational form used for scoring and reporting for the 2022 administration. The educators represented different regions of Kansas and were diverse in gender, race, and urban or rural location. These 15 educators were divided into three panels with five panelists per panel for grades 5, 8, and 11 science assessment. The grade-5 panel evaluated 35 items, and the panels for grades 8 and 11 evaluated 40 items. Before the workshop, a content expert from EdMetric independently assigned alignment ratings to assessment items and standards.

The workshop started with training in alignment for panelists; panelists then acknowledged they understood the process and their roles before they started rating. Two panelists indicated they still needed additional training to understand the alignment process. EdMetric staff met with them individually to help. When independent ratings started, the panelists reviewed the alignments determined by EdMetric's content expert and made changes as needed. Panelists then discussed any disagreement on alignment ratings after individual ratings. Next, panelists rerated disagreed-upon items and discussed any disagreement after rerating. For the whole alignment study, panelists indicated that they could provide rational for their alignments (Egan & Davidson, 2022b).

EdMetric used the modified Webb approach (Webb, 1997, 1999) to evaluate alignment of items to content standards and the blueprint. The evaluation again included the four criteria of categorical concurrence, DOK, ROK, and BOR. EdMetric also evaluated the dimensions measured by science items. The next section summarizes alignment results for both content standards and blueprint according to these four criteria as well as a multidimensionality evaluation.

Among 115 science items, panelists rated three grade-5 items, four grade-8 items, and seven grade-11 items aligning to off-grade standards. Those 14 items rated aligned to off-grade standards are not included in the analysis for these criteria.

III.1.3.1. Categorical Concurrence for Science

Categorical concurrence refers to the degree of similarity and consistency between the standards and assessment content. Categorical-concurrence criterion evaluates (a) the number of items per domain (for alignment between assessment and content standards), and (b) the differences of item distribution between the assessment and the blueprint (for alignment between assessment and blueprint).

If one domain has six or more items, the alignment between assessment and content standard is strong. The alignment becomes weaker as the number of items decreases: five or more items means moderate alignment, and four or more items means weak alignment. Results indicated all domains of three grades, except grade-11 engineering, have strong alignment with the assessments; grade-11 engineering has weak alignment with the assessment.

For each claim, strong alignment between assessment and blueprint means differences between the expected blueprint percentage of items and the actual percentage of items based on panel ratings are smaller than 5%, moderate alignment means differences are between 5% and 10%, weak alignment means differences are between 10% and 15%, and no alignment are for differences larger than 15%. For most claims across the three grades, alignments between blueprint and assessment are strong or moderate. However, for Earth and space science in grade 5, the alignment is weak, and the actual percentage is higher than the expected percentage. For Earth and space science and physical science in grade 11, there is weak alignment and no alignment between blueprint and assessment. The actual percentage of Earth and space science items is lower than the expected range and the actual percentage of physical science items is higher than the expected range.

III.1.3.2. Depth of Knowledge for Science

DOK evaluates alignment of cognitive complexity between assessments and standards. The content expert from EdMetric gave a target DOK range of Level 3 for each standard in the 2013 Kansas Standards for science. The blueprint also specifies the DOK goals as Level 2–3 for all items on science assessments.

Panelists rated the DOK levels of each item. EdMetric compared these item-level DOK ratings with the content-standard DOK target ranges and blueprint DOK goals. The DOK criterion evaluates (a) the percentage of items per domain at or above the target DOK ranges (for alignment between assessment and content standards), and (b) the differences in DOK distribution between the assessment and the blueprint (for alignment between assessment and blueprint).

Results indicated that all items had strong alignment between DOK goals specified by the blueprint and item-level DOK, with more than 85% items rated at DOK Level 2 or 3 by the panelists. However, items were either weakly aligned or not aligned to the content-standard target DOK, with less than 40% items rated at DOK 3 by the panelists.

III.1.3.3. Range of Knowledge for Science

ROK evaluates the extent to which the assessment covers the standards (Webb, 1997). This evaluation criterion examines the percentage of targets measured by at least one item in each domain for the content standards and in each claim for the blueprint.

For each domain and claim of the three science grades, results indicated that both the content standards and blueprint had strong alignment with the assessment, according to the ROK evaluation criteria; more than 67% of targets were measured by at least one item in each domain for the content standards and in each claim for the blueprint.

III.1.3.4. Balance of Representation for Science

BOR examines how items are distributed across the standards. This alignment criterion examines whether items in a domain and claim are evenly distributed across the targets within that domain for the content standards and within that claim for the blueprint.

The results indicated that physical science of grade-5 science was weakly aligned, and all other domains in all grades were strongly aligned between the assessment and the content standards for BOR. For BOR alignment between blueprint and assessment, all claims were strongly or moderately aligned.

III.1.3.5. Multidimensionality for Science

The Kansas Standards for Science require items to measure multiple dimensions because of the three-dimensional nature of Kansas Standards for Science. The multidimensionality of assessments are evaluated by the percentage of items measuring two or more dimensions as determined by the panelists.

The results indicated 63% of grade-5 items, 58% of grade-8 items, and 40% of grade-11 items measured at least two dimensions. According to EdMetric criteria, assessments in grades 5 and 8 have strong alignment and the grade-11 assessment has moderate alignment in the multidimensionality evaluation.

In summary, the alignment between assessment and content standards is strong for most domains across grades on categorical concurrence, ROK, and BOR. For all domains across grades, the alignment between assessment and content standards is weak or not aligned on DOK. Moreover, assessment and blueprint are strongly or moderately aligned for categorical concurrence, DOK, ROK, and BOR for all claims except the Earth and space science claim in grades 5 and 11, and the physical science claim in grade 11. For these three claims, there is either weak or no alignment between assessment and blueprint on categorical concurrence. Finally, all science assessments had strong or moderate alignment on the multidimensionality evaluation.

III.1.3.6. AAI Response to Science Alignment

EdMetric (Egan & Davidson, 2022b) noted that weak or no alignment of DOK between assessment and content standards was caused by a discrepancy between some content standards; a content expert from EdMetric rated the standards at DOK Level 3, while alignment panelists rated their items at DOK Level 2. However, our blueprint requires items be at DOK Level 2 or 3, so it is reasonable that some items on the assessment were at DOK Level 2 to meet the blueprint requirement. We agree with EdMetric that this alignment finding is not surprising (Egan & Davidson, 2022b).

Also, the KAP science alignment study indicated the Earth and space science claim in grades 5 and 11, as well as the physical science claim in grade 11, have either weak or no alignment between assessment and blueprint on categorical concurrence. The Earth and space science claim in grade 5 and physical science claim in grade 11 are overemphasized according to blueprint

distribution. However, the Earth and space science claim in grade 11 does not have enough items because six of 10 Earth and space science items rated as aligning to off-grade standards. We will conduct an internal review of these six items as well as of content standards. We will then work with KSDE to determine if additional items are needed for grade-11 science.

III.2. Validity Evidence Based on Response Process

Response-process evidence examines the extent to which the cognitive skills and processes that students use to answer an item match those targeted by item writers. The evidence is established during the item-development process and with the development of performance level descriptors (PLDs).

Webb's (1997) DOK model is used to identify the cognitive complexity of KAP items, ensuring that items cover the range of cognitive complexity. During the item-development process, items were written by item writers who had been trained on DOK, and item writers either assigned the DOK level to items they wrote or wrote items to reflect the target DOK level. The blueprints imply a target distribution of DOK. The DOK component guided item writers to use language that elicits the cognitive process required by the blueprint and guided item reviewers to evaluate the cognitive process required by items using their experience with students.

PLDs reflect the cognitive processes required for specific content areas. Policy PLDs are the same across grades and subjects and provide the introductory statement for each performance level.

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

As performance levels rise, the expectations of students' proficiency or cognitive processes also rise. As shown across levels 1 through 4, above, the required ability of students to understand and use skills and knowledge changes from limited to basic to effective to excellent.

Moreover, the grade-specific PLDs describe what students know and can do at each performance level. Appendix C includes grade-specific PLDs for all subjects and grades. These grade-specific PLDs provide more-detailed statements for each performance level. For example, as the performance level of grade-5 science rises, the required cognitive processes become more complex. For Level 2, students only need to be able to use a model to describe that matter is made of particles too small to be seen. For Level 3, students need to develop a model to describe that matter is made of particles too small to be seen. For Level 4, students need to develop models to explain different types of matter made of particles too small to be seen.

III.3. Validity Evidence Based on Internal Structure

As described in the *Standards* (APA 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). Three sets of validity evidence about internal structure provide evidence that (a) the KAP assessment is essentially unidimensional, (b) the item response theory (IRT) model used for each subject showed good fit results, and (c) the test contains no or few items flagged for significant and large differential item functioning (DIF), which helps support comparable measurement across groups.

III.3.1. Dimensionality

We applied confirmatory factor analysis (CFA) to evaluate whether a model with one dominant dimension fit the data reasonably well when the IRT scale was set. We carried out CFA using tetrachoric or polychoric correlations for binary or 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 .95 or greater and the Root Mean Square Error of Approximation (RMSEA) was .05 or smaller.

For ELA and mathematics in grades 3–8, CFI ranged from .96 to 1.0, TLI ranged from .96 to 1.0, and RMSEA ranged from .01 to .03. For grade-10 mathematics, both the CFI and the TLI were around .98 and the RMSEA was .02. Thus, the grade-10 mathematics test may be reasonably treated as unidimensional. Overall, for science tests in grades 5, 8, and 11, both the CFI and the TLI were around .99 and the RMSEA ranged from .01 to .03. All tests may be reasonably treated as unidimensional.

III.3.2. Item Response Theory and Model Assumptions

We analyzed KAP items 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, IRT has several model assumptions that need to be fulfilled: model fit, local independence, and item-parameter invariance. The resulting inferences from any application of IRT depend on the degree to which the underlying assumptions are met.

This section describes the IRT models and calibration procedures used for all subjects and grades. The evaluation of IRT assumptions is presented as evidence of score validity. The evaluation analyses of the IRT calibration and assumption occurred when the IRT scale was set. For ELA and mathematics in grades 3–8, all analyses occurred in 2015. For grade-10 mathematics, all analyses occurred in 2022. For science, all analyses occurred in 2017.

III.3.2.1. Item Response Theory Calibration

We used IRT to calibrate item parameters to create a scale for each subject and grade. The IRT scale was set in the first year of operational administration. The next subsections introduce the IRT models, the sample used for calibration, the psychometric software, and the calibration procedures used for KAP.

III.3.2.1.1. Item Response Theory Model

We applied the two-parameter logistic (2PL) model (Birnbaum, 1968) and the graded response model (GRM; Samejima, 1969) to dichotomous and polytomous scored items, respectively. The choice of these two models contributed to the consistent and coherent interpretation of item parameters, as the 2PL model 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* indicates how well the item distinguishes between students with higher or lower levels of proficiency; *difficulty* indicates how hard an item is 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 .5.

III.3.2.1.2. Sample

We cleaned the student data file before calibration. The estimation sample included all students who completed at least five items per test session and exited the test, except for students who needed certain accommodations. Omitted items appeared on the test, but students did not answer them; thus, these omitted items were scored as incorrect answers (coded as 0). Table III-1 provides the number of students by subject and grade for the sample used in IRT scale-setting calibration.

Table III-1. Year, Sample Size, and Number of Items for Scale-Setting Calibration by Subject and Grade

Subject	Year	Grade	Sample size	No. of items
English language arts	2015	3	33,227	327
		4	32,424	310
		5	32,976	313
		6	33,088	320
		7	32,612	304
		8	33,659	293
		10	33,146	315
Mathematics	2015	3	33,197	235
		4	32,391	255

		5	32,805	237
		6	33,070	205
		7	32,609	225
		8	33,725	235
	2022	10	32,378	65
Science	2017	5	33,156	50
		8	33,458	60
		11	32,210	59

III.3.2.1.3. Software

The mirt package (Chalmers, 2012) in R was used for IRT model estimation. The item-parameter calibration used the expectation–maximization algorithm. For all subjects and grades, the IRT calibrations converged; that is, the log-likelihood changes were smaller than 0.0001.

III.3.2.1.4. Calibration Procedures

ELA items and mathematics items for grades 3–8 were administered as operational field-test items in 2015, and 2015 test-administration data were used to set the IRT scale and estimate item parameters. Grade-10 mathematics items were administered as operational field-test items in 2022, and item parameters were estimated using the 2022 test-administration data. Science items were field tested in 2016 and were administered as a fully operational test for the first time in 2017. Science item parameters were estimated using 2017 test-administration data. For each subject and grade, a single-group concurrent calibration was conducted to place all item parameters onto the same scale for each subject and grade assessment. To accomplish this, we compiled all operational items of the same subject and grade into one file to create a student-by-item data matrix, which was then analyzed using estimation software for calibration. The sample size and number of items calibrated are in Table III-1.

III.3.2.2. IRT Model Evaluation

The validity inferences from the IRT results depend on the degree to which assumptions of the models are met and how well the models fit the data. This section describes how the assumptions on IRT model fit, local independence, and item-parameter invariance are evaluated. All operational field-test items were included in model evaluations for ELA and mathematics in grades 3–8. Only items retained for operational scoring were included in model evaluations for grade-10 mathematics and science.

III.3.2.2.1. Model Fit

We used the marginal χ^2 fit statistic to evaluate the model fit for individual items for ELA, mathematics in grades 3–8, and science. 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 of .05, less than 20% of items for ELA and for mathematics in grades 3–8 were flagged as misfit. The *2015 KAP Technical Manual* includes detailed information about the percentage of items flagged as misfit by grade. Using the same significance level, no science items in grades 5 and 8 were flagged as misfit and only four grade-11 items were flagged as misfit.

For grade-10 mathematics, due to the change of calibration software, we used the Q1 chi-squared (χ^2) fit statistic to evaluate the model fit for individual items. We computed the statistics using the mirt package in R during item calibration. The Q1 χ^2 fit statistic followed the χ^2 distribution with degrees of freedom (*df*) equal to the number of possible total scores minus 1. Because the χ^2 tests are sensitive to sample size, we also used the effect size to evaluate item fit. The effect size for χ^2 tests was calculated using Cramér's V (Cramér, 1946). A small Cramér's V effect size is between $0.1/\sqrt{df}$ and $0.3/\sqrt{df}$. A medium Cramér's V effect size is between $0.3/\sqrt{df}$ and $0.5/\sqrt{df}$. A large Cramér's V effect size is greater than $0.5/\sqrt{df}$ (Cohen, 1992). Items whose χ^2 tests were significant at α level of .01 and exhibited a medium to large effect size were flagged

for model-fit issues. For all 56 grade-10 mathematics items, no items were flagged for model-fit issues.

III.3.2.2.2. 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. When student responses to items in latter positions of the test depend on the student responses to their predecessors, then the dependence violates local independence. In this case, when students answer the first item of the group 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.

For all subjects and grade assessments, we used the chi-squared (χ^2)-based local dependence (LD) statistic (Chen & Thissen, 1997) to detect the item pairs with LD. The χ^2 LD index of one item pair followed the χ^2 distribution with degrees of freedom (*df*) equal to 1. Because the χ^2 tests are sensitive to sample size, we also used the effect size to evaluate item fit. The effect size for χ^2 tests was calculated using Cramér's V (Cramér, 1946). For ELA items and for mathematics items in grades 3–8, less than 0.1% pair of items was detected with medium effect-size LD and no pair of items was detected with large effect-size LD. For the grade-10 mathematics assessment and the science assessment of the three grades, no pair of items was detected with a medium or large effect-size LD.

III.3.2.2.3. Parameter Invariance

IRT models claim that item-parameter estimates are invariant up to a linear transformation for all examinees. The strong relationships of item parameters calibrated from two samples indicate that the assumption of parameter invariance is met.

For ELA and mathematics in grades 3–8, we used the Pearson product-moment correlation to evaluate the relationship between the item parameters estimated from two randomly divided student groups for one of the operational forms. To avoid statistical bias caused by outliers, any items with difficulty parameters greater than |6| were excluded from the comparison. For both subjects, the relationships between item-parameter estimates for two samples were strong, with almost all Pearson correlations near 1. The *2015 KAP Technical Manual* includes the correlation values by subject and grade.

For grade-10 mathematics, we used the Pearson product-moment correlation to evaluate the relationship between the item parameters estimated from two randomly divided student groups. These two randomly divided student groups were expected to have the same ability distributions. The correlation for the item-discrimination parameters was .996, and the correlation for the item-difficulty parameters was .998. Both correlations highly supported the parameter-invariance assumptions.

For science, we used the Pearson product-moment correlation to evaluate the relationship between the item parameters estimated from subgroups that were 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|, were excluded from the comparison. The subgroups were determined by gender. The correlations for item-discrimination parameters were .96, .90, and .96 for grades 5, 8, and 11 respectively; the correlations for item-difficulty parameters were .98, .96, and .92 for grades 5, 8, and 11 respectively. In summary, all the Pearson correlations were above .90. These results strongly supported the invariance assumption for KAP science, especially for item-difficulty parameters.

III.3.3. Differential Item Functioning

DIF analysis evaluates items for potential bias and examines whether an item shows statistical difference between two groups of students while controlling for student ability. We used logistic regression to detect items with uniform DIF (i.e., items that are consistently more difficult across all ability levels for one group of students than the other group).

When using the logistic regression for detecting uniform, we predict the probability of a correct response of an item given group and total scale score. The logistic regression equation for each item included a matching variable of the student's total scale score and a group indicator variable. Two logistic regression models were fitted for each item:

$$M_1: logit(P) = \alpha + \beta_1 SS, \qquad (\text{III} - 3)$$

$$M_2: logit(P) = \alpha + \beta_1 SS + \beta_2 G, \qquad (\text{III} - 4)$$

where P is the probability of a correct response to the item, SS is the total scale score, G is the group indicator, α is the intercept, and βs are the slopes. In KAP, there are polytomous items with more than two item score levels and logistic regression only works if there are two score levels. Miller and Spray (1993) suggested switching the group indicator and item score. Thus, the dependent variable is the group indicator and the independent variables are the item score and the scale score in the logistic regression model. This method works greatly for the binary logistic regression model and yields exactly the same results before the switching was made.

The chi-square test of log odds ratio between two logistic regression models listed above (Equation III-3 and III-4) is used to detect the existence of uniform DIF. Because the chi-square test is very sensitive to sample size and KAP items have a large number of students taking them, we used the Jodoin and Gierl (2001) DIF classification criteria to indicate the degree of DIF (i.e., negligible, moderate, large). This classification criteria calculates the Nagelkerke R^2 change first and judges the degree of DIF by the Nagelkerke R^2 change as the effect size. When the DIF test is significant, large DIF is identified by a Nagelkerke R^2 change greater than or equal to .070, moderate DIF has a Nagelkerke R^2 change between .035 and .070, and negligible DIF has a Nagelkerke R^2 change of less than .035.

For each subject and grade, we examined DIF across gender (i.e., female vs. male), race (i.e., Black vs. White), and English learner (EL) status (i.e., EL vs. non-EL). For all subjects and grades, 0 of 831 operational items in the three subjects were flagged for moderate or large gender-related DIF, race-related DIF, or EL-status-related DIF. All results suggested that the item-development process and procedures effectively addressed potential bias-and-sensitivity issues during the development phase.

III.4. Validity Evidence Based on Relations 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" (APA et al., 2014, p. 16). To provide validity evidence based on relations to other variables, we calculated the correlations among different KAP subject scores and compared the KAP and NAEP performance.

III.4.1 Relationships Among KAP Subjects

Past studies showed high correlations between subjects, which indicates that subjects share some common traits; however, the correlations should not be too high. Table III-2 shows the correlations and disattenuated correlations (correcting for measurement errors) between subjects of the same grade, with values that range from .68 to .77 for correlations, and from .76 to .87 for disattenuated correlations. The lowest correlations among subjects are between grade-8 ELA and mathematics and between grade-10 ELA and mathematics. The highest correlations are between grade-3 ELA and mathematics and grade-5 ELA and science. After correcting for measurement error, the lowest disattenuated correlation is still between grade-8 ELA and mathematics and grade-10 ELA and mathematics, and the highest disattenuated correlation is between grade-5 ELA and science. According to Cohen (1988), a correlation larger than .50 is considered a correlation with large effect size. All correlations among KAP subjects have large effect size, indicating that some common traits are shared across KAP subjects.

Table III-2. Correlations (C) and Disattenuated Correlations (DC) Among English Language Arts (ELA), Mathematics, and Science Scores

Grade	ELA vs. mathematics		ELA vs. science		Mathematics vs. science	
	C	DC	C	DC	C	DC
3	.77	.83	-	-	-	-
4	.73	.80	-	-	-	-
5	.72	.79	.77	.87	.72	.80
6	.74	.82	-	-	-	-
7	.72	.80	-	-	-	-
8	.68	.76	.73	.85	.69	.79
10	.68	.76	-	-	-	-

Note. ELA = English language arts.

III.4.2. Relationships Within a KAP Subject

The correlation between current-year and previous-year KAP scores of one subject for the same students should be high because similar constructs are measured across grades within a subject. Table III-3 shows the correlations and disattenuated correlations (i.e., correcting for measurement errors) between adjacent grades of the same subjects in 2022 and 2021. For the grades in which all students did not take KAP assessments in the previous year, that is, no KAP assessment for the adjacent grade in the previous year, the correlations are not calculated. Values range from .80 to .84 for correlations, and from .89 to .91 for disattenuated correlations. The correlations and disattenuated correlations between grades are very similar for one subject, and

ELA correlations are slightly lower than correlations in mathematics. All correlations between adjacent grades within a subject are very high and have large effect size, indicating that similar constructs are measured within KAP subjects.

Table III-3. Correlations (C) and Disattenuated Correlations (DC) Between Adjacent Grades for English Language Arts and Mathematics

Grade	English language arts		Mathematics	
	C	D	C	D
4 vs. 3	.82	.90	.83	.89
5 vs. 4	.81	.90	.84	.90
6 vs. 5	.80	.89	.83	.90
7 vs. 6	.80	.90	.84	.91
8 vs. 7	.80	.89	.81	.89

III.4.3. Relationships Between KAP Assessment and National Assessment of Educational Progress

The state of Kansas participates in the NAEP, also known as the Nation's Report Card. NAEP is the largest nationally representative assessment of what American students know and can do, and it serves a different role than state assessments do. NAEP assessments allow each state to be compared to national results and to evaluate progress over time. The results inform the public about the academic achievement of elementary (grade 4) and secondary (grade 8) students in Kansas and in the United States in ELA and mathematics.

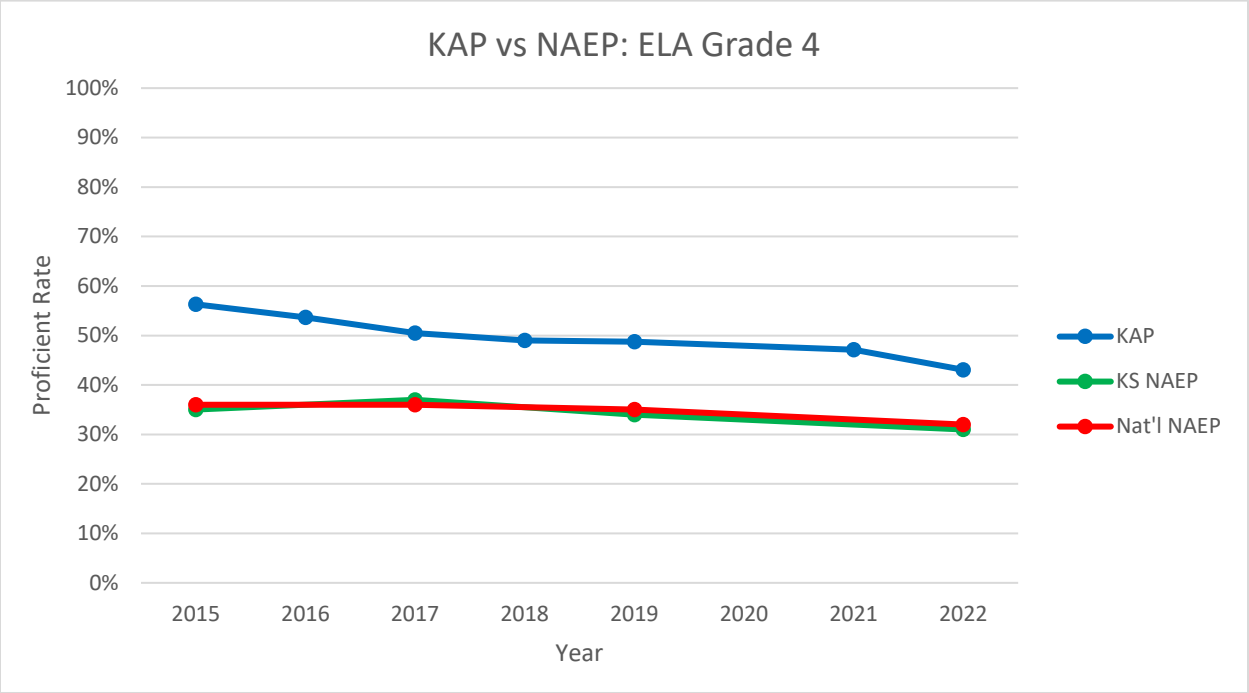
Thus, the relationship between KAP and NAEP performance is expected to be strong. Because individual NAEP scores are not available, only the trend of proficiency rates across years is compared between the two assessments. KAP and NAEP assessments use different achievement standards to judge whether a student meets proficiency. Comparing proficiency rates within a year is not as meaningful as comparing trends of proficiency rates across years. The trends of the two assessments can indicate the actual performance of Kansas students based on the two assessments measuring a similar construct. KSDE provides more information about NAEP on the KSDE website.

KAP categorizes student performance by four performance levels (i.e., 1, 2, 3, 4). The proficiency rate of KAP is the percentage of students in levels 3 and 4. NAEP categorizes student performance by three performance levels (Basic, Proficient, and Advanced). The proficiency rate of NAEP is the percentage of students in Proficient and Advanced levels. Figure III-1, Figure III-2, Figure III-3 and Figure III-4 compare KAP and NAEP proficiency rates across years for ELA and mathematics in grades 4 and 8 from 2015 to 2022. [2] In years 2015 through 2022, KAP proficiency rates ranged from 43% to 56% for grade-4 ELA, from 21% to 32% for grade-8 ELA, from 34% to 40% for grade-4 mathematics, and from 21% to 27% for

[2] NAEP is administrated in odd-numbered years only. The planned 2021 NAEP assessment was delayed to 2022.

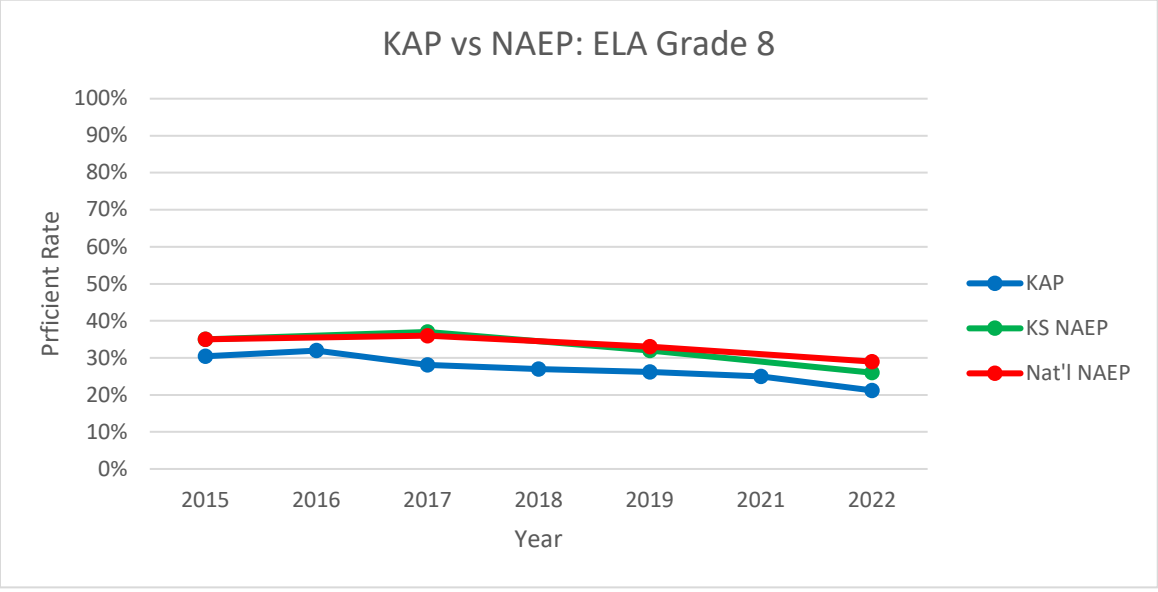
grade-8 mathematics. The Kansas and national NAEP proficiency rates for both ELA and mathematics grades 4 and 8 are very similar across years from 2015 to 2019, ranging from 30% to 40%, with most around 35%. However, there is a decrease in both Kansas and national NAEP proficiency rates for both ELA and mathematics grades 4 and 8 in 2022, with nearly a 5% decrease in grade 4 and close to a 10% decrease in grade 8. The 2022 ELA and mathematics proficiency rates for grades 4 and 8 for KAP, Kansas NAEP, and national NEAP are lower than the 2019 proficiency rates because of the impact of COVID-19. The similar trend of proficiency rates proficiency rates among KAP, Kansas NAEP, and national NAEP shows that performance on NAEP is not different from that on KAP.

Figure III-1. Grade-4 English Language Arts (ELA) Proficiency-Rate Trend Across Years: KAP vs. NAEP



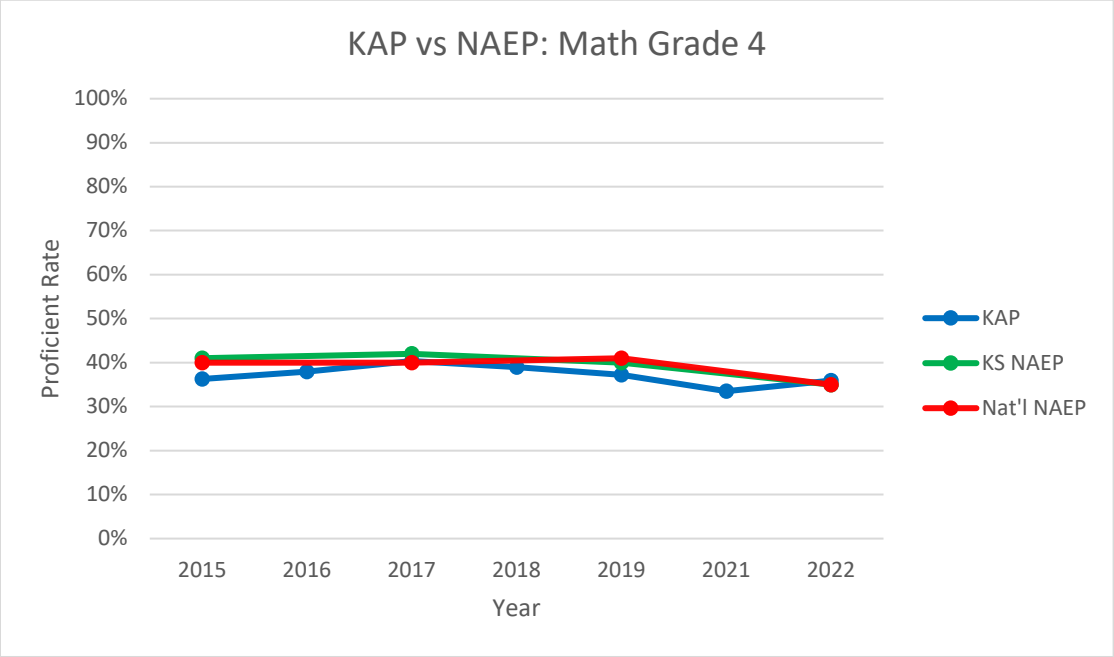
Note. KAP = Kansas Assessment Program; NAEP = National Assessment of Educational Progress.

Figure III-2. Grade-8 English Language Arts (ELA) Proficiency-Rate Trend Across Years: KAP vs. NAEP



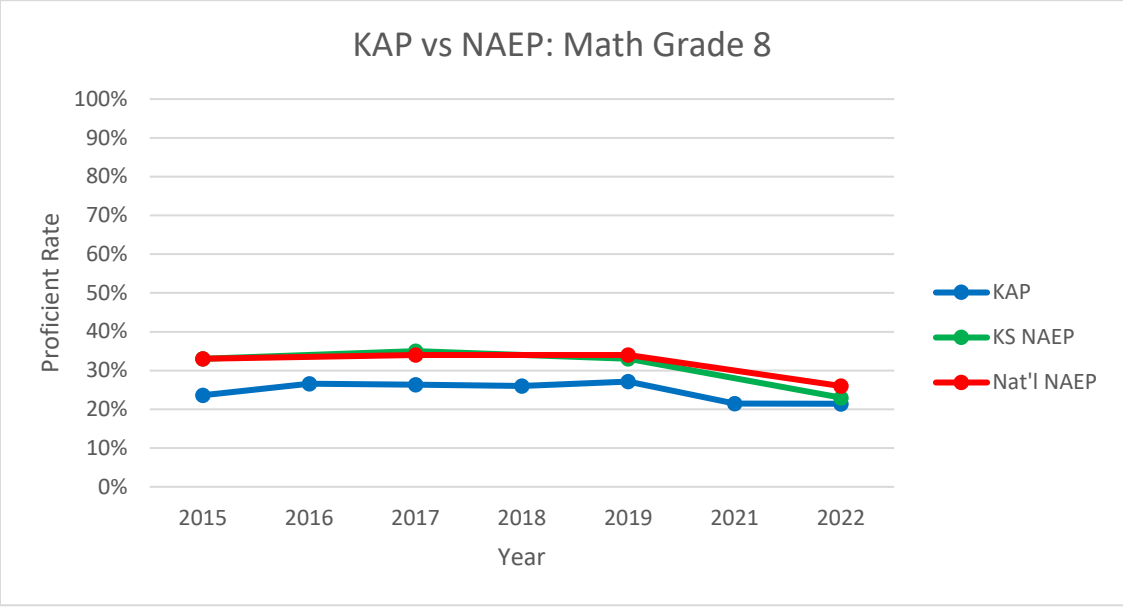
Note. KAP = Kansas Assessment Program; NAEP = National Assessment of Educational Progress.

Figure III-3. Grade-4 Mathematics Proficiency-Rate Trend Across Years: KAP vs. NAEP



Note. KAP = Kansas Assessment Program; NAEP = National Assessment of Educational Progress.

Figure III-4. Grade-8 Mathematics Proficiency-Rate Trend Across Years: KAP vs. NAEP



Note. KAP = Kansas Assessment Program; NAEP = National Assessment of Educational Progress.

III.5. Validity Evidence Based on Consequences of Testing

Validity evidence based on consequences refers to evidence supporting the intended uses and interpretation of test scores. A primary intended use of KAP test scores is to provide scores that can be used with local assessment scores to assist in improving a building's or district's programs as stated in the *Kansas Assessment Examiner's Manual 2021–2022*. Section IV.4. Scoring and Scaling summarizes how items and tests are scored. For a given test score, the performance level is determined by a set of established cut scores. Chapter VI Academic Achievement Standards and Reporting summarizes the cut scores and includes an example of a KAP student score report. To help educators and parents interpret KAP results, KAP also provides the *KAP Educator Guide* and the *KAP Parent Guide*.

To evaluate how educators use KAP test scores, we collected data in a 2022 KAP teacher survey. Two hundred eighty-two educators, about 1% of all educators in Kansas, responded to the KAP teacher survey. Among the educators who responded, 77% were classroom teachers. A total of 264 ELA, 260 mathematics, and 232 science educators evaluated whether KAP assessment results provide useful information when planning for classroom instruction for the next school year. Of the educators who responded to this question, 58% of ELA educators, 53% of mathematics educators, and 49% of science educators either agreed or strongly agreed that KAP results were useful for planning for instruction. Some educators also described other uses of KAP assessment results in addition to planning for instruction. Those other uses include:

- Teacher professional development
  - KAP results were used to determine teacher needs and professional-development content.
- Grouping students
  - KAP results were used to determine different interventions or group instructions.
- Identifying students at risk
  - KAP results identified students who were at risk, who needed more instruction, or who needed interventions.
- Student placement
  - Students were placed in a different level of mathematics class based on KAP results. Also, KAP results were used for placement in advanced classes and remediation classes.

Moreover, schools and districts implemented year-round, in-person instruction (i.e., no remote or hybrid instruction), but COVID-19 may have continued to affect students' learning and learning experiences; for example, students may have missed instruction because of illness or quarantine. More-detailed information about the effects of COVID-19 on instruction in 2022 is in Section IV.4.3.3.1. Monitoring the COVID-19 Effect. Caveat language in the student score report, KAP Educator Guide, and KAP Parent Guide reminds students, parents, and educators that learning conditions and student performance may have been affected by COVID-19. The caveat states

> When interpreting KAP results, please take into consideration how the conditions for learning, which may have been disrupted by the pandemic, may influence performance. (*KAP Parent Guide*, KSDE, 2022)

Parents and educators can still use test scores to help identify students' relative strengths and limitations, to determine their progress in meeting state curriculum standards, and to compare their performance to that of other students in the school, district, and state, as stated in the *KAP Parent Guide* (KSDE, 2022); however, parents and educators need to consider the impact of the COVID-19 pandemic on learning.

IV. Technical Quality: Other

This chapter mainly describes evidence related to the technical quality of the Kansas Assessment Program (KAP) and summarizes the technical analysis for ongoing maintenance, such as additional monitoring of item responses in 2023 for technology-enhanced items. Most of the analysis described in this chapter is based on 2022 assessment data. Evidence for technical quality includes test reliability, fairness and accessibility, an item-analysis summary, a test-analysis summary, and trend data.

IV.1. Reliability

Reliability is a test-score-consistency index that shows the degree of test-score consistency across repeated measures. Test scores that are stable across repeated measures indicate a more reliable test. Factors leading to unstable test scores are called *measurement errors*. Measurement errors include, but are not limited to, changes in testing conditions; changes in a student's knowledge, physical condition, or mental status; and changes in testing content across multiple test administrations. Measurement errors cannot be fully removed but can be reduced. For example, standardized testing procedures reduce measurement errors caused by changing testing conditions. KAP has standardized testing procedures, and the same procedures are applied to all students; specific accommodations are provided to students with special needs. The *Kansas Assessment Examiner's Manual 2021–2022* describes these testing-procedure specifications.

In the context of educational achievement tests, factors such as learning, fatigue, and motivation may affect test takers at different rates for repeated measures. It is impractical to test the same content area repeatedly as test takers cannot maintain the same knowledge, physical condition, and mental status across test administrations. Therefore, the reliability for educational measures is typically estimated rather than calculated directly. Estimated reliability coefficients range from 0 to 1. Higher values indicate more-reliable tests with less measurement error.

In this section, we present reliability estimates for overall scores and subscores provided by the KAP assessments. The overall score-reliability estimates are calculated for the full sample of tested students as well as for student groups. We also include item response theory (IRT) information functions and conditional standard errors of measurement at each cut score, as well as classification consistency and accuracy estimates for overall scores. Finally, reliability, classification consistency, and accuracy estimates for KAP subscores are summarized.

IV.1.1. Test Reliability

We used a marginal-reliability method (Green et al., 1984) to estimate test reliability. This method can estimate reliability for both fixed-form and adaptive tests. The calculation formula for marginal reliability is

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

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

As shown in Table IV-1, mathematics reliability estimated by the marginal-reliability method is above .91. Reliability estimates for English language arts (ELA) are above .88. Science has relatively lower reliability estimates because there are fewer test items (35 items for grades 5 and 8, 40 items for grade 11) compared to ELA (47 items) and mathematics (55 items for grades 5–8, 56 items for grade 10), but values are still greater than or equal to .83.

Table IV-1. Test-Reliability Estimate by Subject and Grade

Grade	English language arts	Mathematics	Science
3	.91	.94	
4	.89	.94	
5	.89	.93	.87
6	.89	.93	
7	.88	.93	
8	.89	.92	.83
High school	.88	.91	.87

IV.1.1.1. Student-Group Reliability

We estimated reliabilities by the marginal-reliability method for gender groups, race groups, ethnicity groups, English learner (EL) status groups, and disability status groups. [3] Table IV-2, Table IV-3, and Table IV-4 present student-group reliability estimates for ELA, mathematics, and science. For ELA and mathematics, the reliabilities estimated for each group by the marginal-reliability method were close to or above .90 across grades, ranging from .86 to .92 for ELA, and from .88 to .94 for mathematics. Science had relatively lower subgroup-reliability estimates because the subject had fewer test items compared to ELA and mathematics. Science subgroup-reliability estimates ranged from .83 to .89 across grades. For all three subjects, the differences in reliability estimates among different student groups were small.

[3] Economically disadvantaged status is not shared with ATLAS to protect the privacy of students, so this student group is not included in the comparison.

Table IV-2. Student-Group Reliability Estimate for English Language Arts

Subgroup	Grade						
	3	4	5	6	7	8	10
Gender							
Male	.91	.90	.89	.89	.88	.89	.88
Female	.91	.89	.89	.89	.88	.89	.88
Race							
NA	.92	.91	.90	.90	.89	.90	.89
Asian	.90	.88	.87	.87	.86	.87	.86
Black	.92	.91	.90	.90	.89	.90	.89
NHPI	.92	.90	.89	.90	.89	.90	.89
Other	.91	.90	.89	.89	.88	.89	.88
White	.91	.89	.89	.89	.88	.89	.88
Hispanic							
Yes	.92	.91	.90	.90	.89	.90	.89
No	.91	.89	.89	.89	.88	.89	.88
SWD							
Yes	.92	.91	.89	.90	.89	.90	.88
No	.91	.89	.89	.89	.88	.89	.88
EL							
Yes	.92	.91	.90	.90	.89	.90	.89
No	.91	.89	.89	.89	.88	.89	.88

Note. NA = Native American; NHPI = Native Hawaiian and Pacific Islander; SWD = student with disability; EL = English learner.

Table IV-3. Student-Group Reliability Estimate for Mathematics

Subgroup	Grade						
	3	4	5	6	7	8	10
Gender							
Male	.94	.94	.92	.93	.92	.91	.91
Female	.94	.94	.93	.93	.93	.92	.92
Race							
NA	.94	.94	.94	.93	.93	.92	.92
Asian	.93	.93	.89	.91	.90	.90	.88
Black	.94	.94	.93	.93	.93	.91	.91
NHPI	.94	.94	.94	.94	.93	.92	.91
Other	.94	.94	.93	.93	.93	.92	.91
White	.94	.94	.93	.93	.93	.92	.92
Hispanic							
Yes	.94	.94	.94	.93	.93	.92	.92
No	.94	.94	.93	.93	.92	.92	.91
SWD							
Yes	.94	.94	.93	.93	.92	.91	.91
No	.94	.94	.93	.93	.93	.92	.91
EL							
Yes	.94	.94	.94	.93	.93	.91	.91
No	.94	.94	.93	.93	.93	.92	.91

Note. NA = Native American; NHPI = Native Hawaiian and Pacific Islander; SWD = student with disability; EL = English learner.

Table IV-4. Student-Group Reliability Estimate for Science

Subgroup	Grade		
	5	8	11
Gender			
Male	.86	.83	.87
Female	.87	.84	.88
Race			
Native American	.88	.84	.89
Asian	.84	.83	.85
Black	.88	.84	.89
NHPI	.88	.83	.89
Other	.87	.84	.88
White	.87	.83	.87
Hispanic			
Yes	.88	.84	.89
No	.86	.83	.87
Student with disability			
Yes	.88	.83	.89
No	.86	.84	.87
English learner			
Yes	.88	.83	.89
No	.86	.83	.87

Note. NHPI = Native Hawaiian and Pacific Islander.

IV.1.2. Test Information

For KAP tests, we use IRT models to estimate students' latent ability (theta), which is then transformed to a scale score. Using IRT models, we can estimate test information functions (TIFs) for each theta value across the whole performance continuum. A TIF is computed as the sum of item information function of all operational items in a grade for each test. We use the TIF to estimate the amount of information the test provides at each theta; the TIF is conceptually parallel to the reliability coefficient in classical test theory. Figure IV-1, Figure IV-2, and Figure IV-3 present the TIFs for theta values ranging from −3 to 3 in increments of 0.5 for each grade in ELA, mathematics, and science. The graph also indicates the level-3 theta cuts, which are the proficiency cuts.

Typically, TIF values are high at the center of the theta distribution and gradually decrease toward the two ends of the theta scale, where thetas are very low or very high; this distribution results in a bell-shaped pattern. For ELA, grades 3, 6, and 7 have TIFs reaching the maximum value at −1 theta value and other grades had TIFs reaching the maximum value at −0.5 theta value. The level-3 theta cuts for ELA range from −0.5 to 0.5 across grades. The theta values with maximum TIFs are close to the level-3 theta cuts. Mathematics has TIFs reaching the maximum value at theta value 0 except grade 3, with TIF reaching the maximum value at theta value −0.5. The level-3 theta cuts for mathematics range from −0.7 to 0.7 across grades, with most grades' level-3 theta cuts between 0 and 0.5. Science has TIFs reaching the maximum value at theta

values −0.5 for all grades. The level-3 theta cuts for science range from 0.0 to 0.5 across grades. Among the three subjects, mathematics had the smallest difference between theta value with maximum TIFs and the level-3 theta cut.

Figure IV-1. Test Information Function for English Language Arts

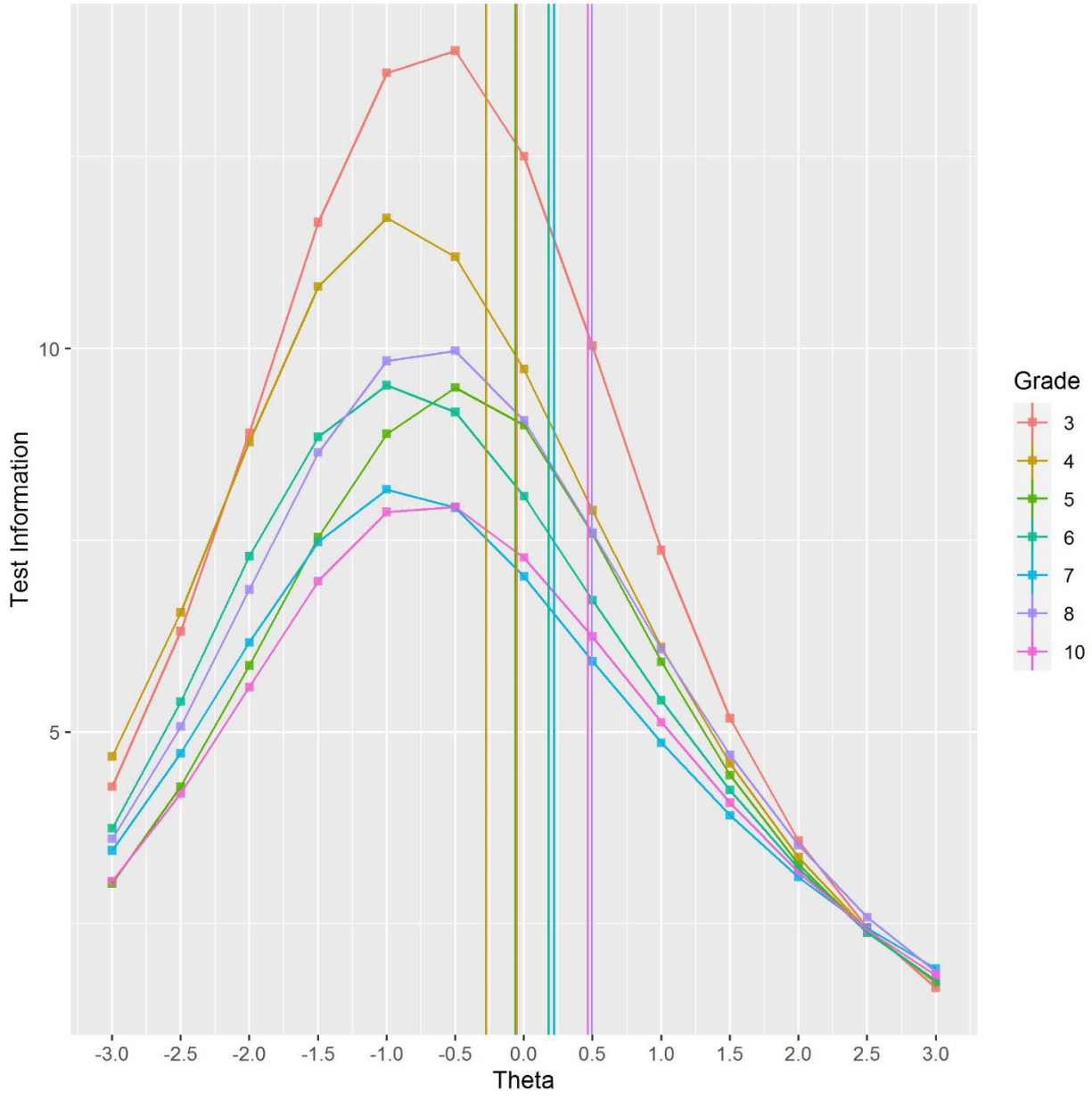


Figure IV-2. Test Information Function for Mathematics

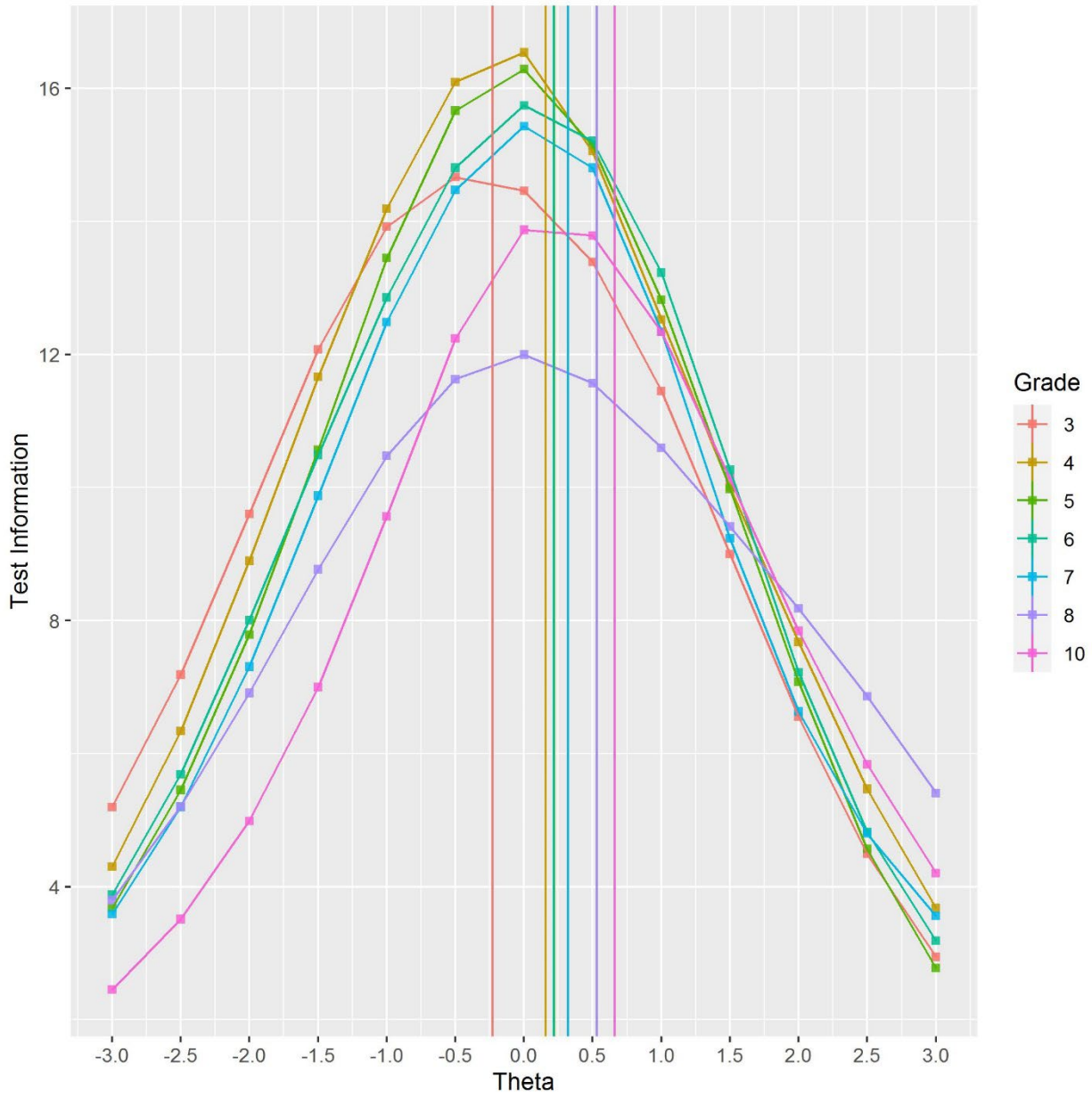
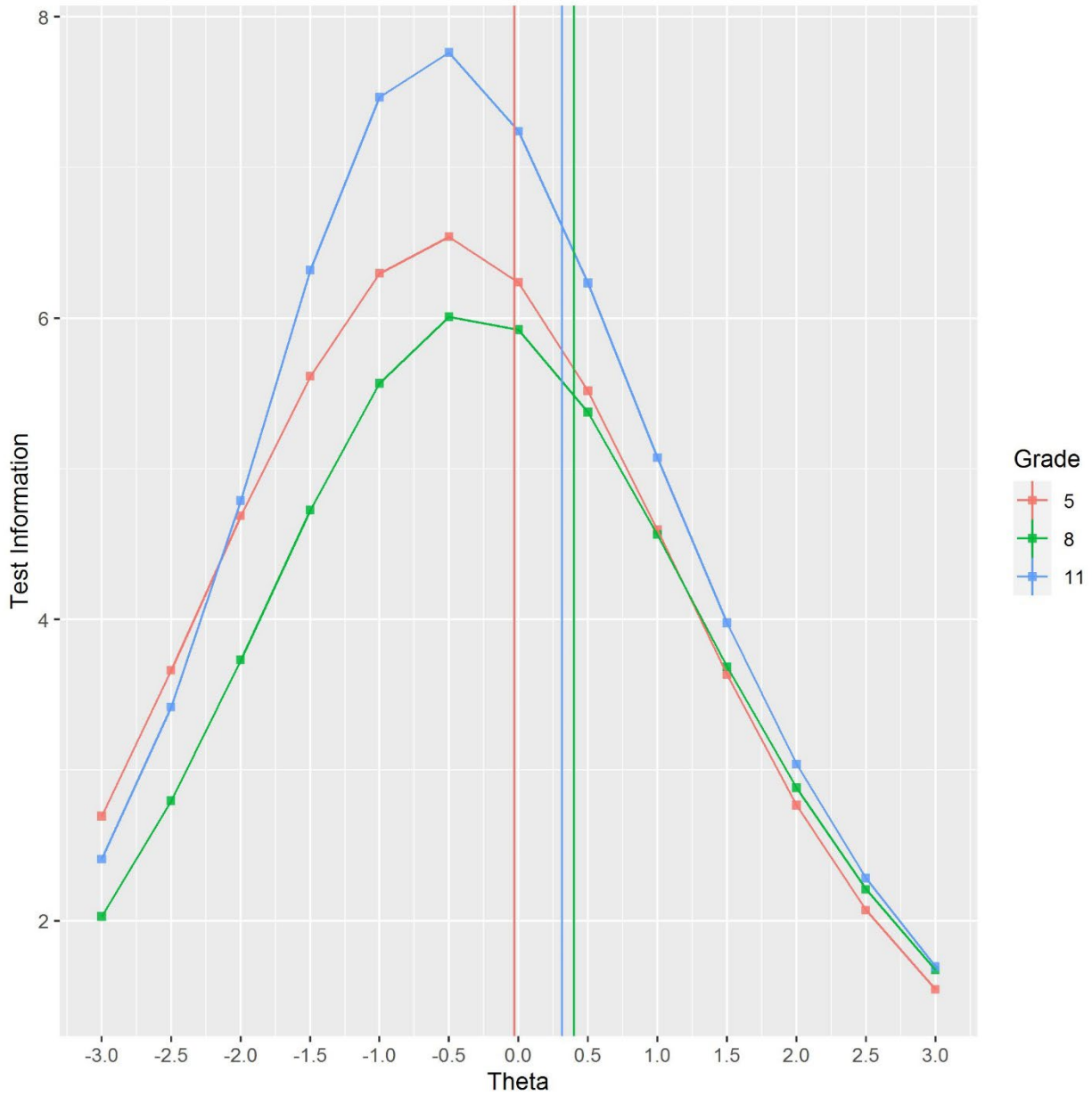


Figure IV-3. Test Information Function for Science



In IRT, we estimate a standard error for each value of theta, called the *conditional standard error of measurement* (CSEM). CSEMs are computed through their inverse relationship with TIFs. For reporting purposes, the CSEM is put on the scale-score metric and reported. The CSEMs at cut scores for levels 2, 3, and 4 of each subject and grade are in Table IV-5.

For ELA and science, level-2 cuts have the lowest CSEMs and level-4 cuts have the highest CSEMs, except for grade 5 in ELA, where the level-3 cut has the lowest CSEM and the level-4 cut has the highest CSEM. For mathematics, level-3 cuts have the lowest CSEMs and level-4 cuts have the highest CSEMs, except for grade 8, where the level-2 cut has the lowest CSEM and the level-4 cut has the highest CSEM. When comparing CSEMs among subjects, we found that

mathematics has the lowest CSEMs at cut scores, and science has the highest CSEMs at cut scores. This pattern is consistent with the marginal-reliability estimates, in which mathematics has the highest marginal reliability and science has the lowest marginal reliability.

Table IV-5. Conditional Standard Error of Measurement at Cut Scores

Grade	English language arts			Mathematics			Science		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
3	6.9	7.2	9.4	6.8	6.5	7.3			
4	7.8	7.9	10.9	6.9	6.2	7.8			
5	8.4	8.3	10.4	6.6	6.3	7.6	9.9	10.1	12.2
6	8.2	9.2	12.5	6.8	6.3	7.6			
7	8.7	9.8	13.0	7.2	6.5	10.1			
8	7.7	9.2	13.9	7.3	7.4	8.8	10.2	10.8	13.1
10	8.8	10.0	13.4	7.0	6.8	8.5			
11							9.0	9.9	12.4

IV.1.3. Classification Consistency and Accuracy

Classification consistency and accuracy indicate how accurately students are classified into performance levels. Performance-level classification consistency and accuracy are of great interest for testing programs that serve as accountability purposes. According to Livingston and Lewis (1995), *classification consistency* refers to "the agreement between the classifications based on two nonoverlapping, equally difficult forms of the test" (p. 180), and *classification accuracy* refers to "the extent to which the actual classifications of test takers on the basis of their single-form scores agree with those that would be made on the basis of their true scores, if their true scores could somehow be known" (p. 180). Both classification consistency and accuracy indices range from 0 to 1, with 0 representing classifications that are not consistent or accurate and 1 representing perfectly consistent or accurate classifications.

Because of the unobservable nature of true scores and the impossibility of repeated testing, actual observed-score distribution and reliabilities are used to estimate a true-score distribution and an observed-score distribution for an alternate parallel form (Livingston & Lewis, 1995). Classification consistency is calculated as the classification agreement between two observed-score distributions (i.e., the observed-score distributions of actual and alternate parallel forms). Kappa is used to calculate the degree of agreement. Classification accuracy is calculated as the probability of accurate classification between the true-score and actual observed-score distributions.

Table IV-6 presents the results for overall classification consistency and accuracy across all four performance levels, as well as for the dichotomies created by the three cut scores. For the overall KAP classification, classification-consistency indices range from .47 to .64, and classification-accuracy indices range from .71 to .83 across all grades and subjects. Classification consistency and accuracy for the KAP level-3 performance-level cut (i.e., 1, 2 vs. 3, 4) is most important because the level-3 cut is the proficiency-level cut. Classification-consistency indices range from .55 to .81, and classification-accuracy indices range from .87 to .99 across all cuts, grades,

and subjects. For all subjects and grades except grade-10 ELA, the level-3-cut classification-consistency index is higher than the other two cuts' classification-consistency indices. Within the same grade, classification consistency and accuracy for the science tests are lower than for the other two subject tests because science tests have fewer items.

Table IV-6. Classification Consistency and Accuracy

Subject and grade	Cut-score category							
	Overall		1 vs. 2, 3, 4		1, 2 vs. 3, 4		1, 2, 3 vs. 4	
	C	A	C	A	C	A	C	A
ELA								
3	.58	.78	.68	.91	.76	.92	.74	.96
4	.55	.77	.58	.91	.72	.90	.69	.96
5	.52	.74	.65	.90	.73	.91	.70	.95
6	.56	.77	.69	.90	.71	.90	.62	.97
7	.54	.77	.68	.89	.70	.91	.62	.97
8	.58	.81	.68	.90	.70	.93	.57	.98
10	.55	.78	.69	.89	.68	.91	.58	.97
Mathematics								
3	.64	.81	.74	.93	.80	.93	.78	.95
4	.64	.82	.65	.92	.81	.94	.78	.97
5	.60	.80	.66	.90	.81	.94	.80	.97
6	.64	.82	.73	.91	.80	.94	.76	.97
7	.60	.81	.55	.88	.81	.95	.76	.99
8	.64	.83	.72	.90	.80	.95	.76	.98
10	.54	.78	.60	.86	.81	.96	.79	.98
Science								
5	.47	.71	.57	.89	.71	.90	.70	.94
8	.47	.73	.62	.87	.67	.91	.61	.96
11	.51	.75	.65	.88	.73	.92	.71	.96

Note. ELA = English language arts; C = consistency; A = accuracy.

IV.1.4. Subscore Reliability

In addition to the total test score, the scores of subsets of ELA, mathematics, and science items are reported as subscores. The number of items in each subscore varies, and some items contribute to multiple subscores. Six is the minimum number of items reported for a subscore.

ELA has a total of six subscores; the same six subscores are reported for all grades (grades 3–8 and 10). The primary subscores are overall reading and overall writing. The overall reading score has two subscores: key ideas and details; and craft, structure, and language in reading. The overall writing score has two subscores: text types and purpose, and language in writing.

The number of mathematics subscores varies across grades. Grade 3 has six subscores, grade 4 has six, grade 5 has five, grade 6 has seven, grade 7 has seven, grade 8 has five, and grade 10 has six subscores. All grades include two separate subscores: skills and concepts, and strategic

thinking and reasoning. Table IV-7 shows the additional subscores for each grade within the skills-and-concepts subscore.

Table IV-7. Subscores for Mathematics by Grade

	Grade						
	3	4	5	6	7	8	10
Skills and concepts	X	X	X	X	X	X	X
1. Operations and algebraic thinking	X	X					
2. Number and operations in base ten		X	X				
3. Number and operations with fractions	X	X	X				
4. Measurement and data	X	X	X				
5. Ratios and proportional relationships				X	X		
6. The number system				X	X		
7. Expressions and equations				X	X	X	
8. Algebra							
9. Functions						X	X
10. Geometry	X			X	X	X	X
11. Statistics and probability				X	X		X
12. Number and quantity and algebra							X
Strategic thinking and reasoning	X	X	X	X	X	X	X

Science has three subscores for each grade: physical science, life science, and Earth and space science. Because the science test is shorter, we did not report additional subscores at a finer grain size for science.

We report these subscores in three categories—below, meets, and exceeds—when comparing them with the performance of level-3 students. When a student responds to less than 60% of the items of a subscore, we report the result as *insufficient data* instead of as a subscore category. We assign subscore categories according to subscore scale scores. The procedure for computing subscore scale scores is similar to that for computing overall test scale scores. We use the summed-score method (Thissen & Wainer, 2001) through IRT models to estimate student latent proficiencies (thetas) in each subscore category, and then linearly transform them to scale scores using the test's scaling constants. We use item parameters derived at the test level to derive thetas for subscores and choose cuts of 300 and 325 (one *SE* above 300) to define students' subscore categories. Subscore scale scores that are less than 300 are categorized as *below*, 300 to 325 are categorized as *meets*, and above 325 are categorized as *exceeds*.

We conducted three analyses to determine the reliability of subscores: reliabilities, classification consistencies, and classification accuracies. Appendix D includes estimates of the marginal reliability, classification consistency, and classification accuracy for different subscores for each subject and grade. In summary, the averages of reliability estimates are approximately .61, .64, and .60 for ELA, mathematics, and science, respectively. The averages of consistency indices are approximately .35, .37, and .34 for ELA, mathematics, and science, respectively. The averages of accuracy indices are approximately .74, .76, and .74 for ELA, mathematics, and science, respectively. The results indicate that the subscores provide reasonable, reliable results. There is some variability in the reliability estimates, classification-consistency indices, and classification-

accuracy indices across each subscore by subject and grade. The subscore-reliability estimates range from .46 to .70 for ELA, from .50 to .80 for mathematics, and from .53 to .65 for science. Classification-consistency indices range from .25 to .48 for ELA, from .19 to .54 for mathematics, and from .28 to .42 for science. Classification-accuracy indices range from .63 to .88 for ELA, from .58 to .88 for mathematics, and from .64 to .81 for science.

The number of items measuring each subscore affects the reliability, classification consistency, and classification accuracy, as we measured some subscores by only six items and other subscores by 47 items. We expect the estimates of reliability, classification consistency, and classification accuracy of subscores with fewer items to be low.

IV.2. Accessibility and Fairness

During the development and administration of the KAP assessment, we considered accessibility for all students and fairness across student groups in every step. We used universal design (UD) as a guide during the development of items, test formats, and the online test-delivery interface to ensure fairness and accessibility for all students. Section IV.2.2. Fairness summarizes the UD guidelines. All operational items pass a bias-and-sensitivity review to mitigate the likelihood of content bias toward any one student group. The bias-and-sensitivity review described in Section II.3.4.2.3. Item Fairness-Review Process has external reviewers review items to identify unfairness barriers that may prevent students from demonstrating what they know and can do.

IV.2.1. Accessibility

According to the *Standards for Educational and Psychological Testing*, "accessibility is the degree to which the items or tasks on a test enable as many test takers as possible to demonstrate their standing on the target construct without being impeded by characteristics of the item that are irrelevant to the construct being measured" (American Psychological Association [APA], 2014, p. 215; hereafter *the Standards*). Evidence in support of accessibility of an assessment comprises inclusion, accommodations, and the implementation of UD in items and test development. UD refers to principles that provide equal access to all students. Section IV.2.2. Fairness summarizes the implementation of UD in item and test development. However, some barriers, such as blindness, cannot be addressed by UD. Test inclusion and accommodation policies help address these needs. The Kite® online test system provides many accommodations, including magnification, text-to-speech, and color contrast, among others. Some students require braille tests, which are made available to students who need them. For more details about accommodations for KAP, see Chapter V. Inclusion of All Students.

The 2022 KAP teacher survey asked teachers about the accessibility supports on KAP. Among the 279 educators (approximately 1% of educators in Kansas) who responded to the question about accessibility supports, 253 (91%) agreed or somewhat agreed that their students had access to all necessary accessibility supports to participate in the assessment. While the results suggest that KAP provides students with necessary accessibility supports, additional data from a larger sample of teachers is needed.

IV.2.2. Fairness

According to the *Standards*, "the central idea of fairness in testing is to identify and remove construct-irrelevant barriers to maximal performance for any examinee" (APA et al., 2014, p. 74). The *Standards* identifies fairness as an issue related to the validity of test-score inferences. Evidence supporting the assertion of fairness in an assessment comes from several stages, such as the item- and test-development stages before test administration and the differential item functioning (DIF) analyses stage after test administration.

Using appropriate item- and test-development processes is an excellent start for ensuring fairness. UD in item and test development not only allows for the participation of the widest range of students, but also bolsters the validity of score inferences. KAP's comprehensive inclusion rules mean that KAP tests include all Kansas students (details about the policy of including all students are in Chapter V. Inclusion of All Students). While the initial intention is to meet the assessment requirements of special-needs students, the benefits of universally designed assessments should apply to all students with diverse characteristics. Item-writer training informs participants about UD concepts, includes a definition of UD, and provides examples of test items that adhere to UD principles. Additionally, item-writer guidelines comprise many UD principles. The following are UD guidelines used during KAP test 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.

For DIF results, see Section III.3.3. Differential Item Functioning. DIF analyses conducted for the current administration indicate that no items were identified with significant DIF across gender (i.e., female vs. male), race (i.e. Black vs. White), and EL status (i.e., EL vs. non-EL) for all three subjects. DIF analysis examines whether an item shows any statistical difference between two groups of students after controlling for student proficiency. No items with DIF contribute to the evidence in support of fairness during item writing and reviewing.

IV.3. Full Performance Continuum

KAP was designed and developed to produce a reasonably precise estimation of student proficiency across the full performance continuum in each subject area and grade. TIFs across different ability levels and conditional error of measurements at the cut scores from Section IV.1.2. Test Information show test precision across the full range of ability estimates. Results indicate that KAP tests can accurately estimate ability across the full theta scale, especially in the middle of the scale.

Another approach to cover the full performance continuum is to use items that cover different cognitive complexity levels and a wide range of difficulties. We measure KAP items' cognitive

complexity levels by the depth of knowledge (DOK) framework (Webb, 1997). The blueprint specifies the expected DOK ranges for each cluster (included in Appendix A). When test items are written to each cluster, the items also have to reflect the expected DOK level as implied by the content to be measured. We emphasize this expectation throughout item writing and during both internal and external item reviews. Consequently, items selected for a test to meet the blueprint also meet the underlying DOK requirements. During test construction, we screen item quality through item difficulty, 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 for measuring student abilities. Additionally, we plot test-characteristic curves, test information, and CSEM during test construction to gauge the proficiency range of each test. To confirm that the tests efficiently cover the full performance continuum as expected, we present classical and IRT item statistics as well as DOK count here as evidence.

IV.3.1. Classical Item Statistics

Here we calculate and provide two statistics: item difficulty and item discrimination. *Item difficulty* refers to the difficulty of an item, 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 answer the item correctly. Equation IV-2 shows the calculation of the *p* value.

$$p \text{ value } = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i}{\text{item max score}}, \qquad \text{(IV-2)}$$

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.

Table IV-8, Table IV-9, and Table IV-10 present summaries of item difficulty for ELA, mathematics, and science tests. The grade average item difficulties range from .50 to .53 for ELA; from .46 to .50 for mathematics; and from .45 to .53 for science. For all grades and subjects, the ranges of item difficulty are large, ranging from .17 to .81 for ELA, from .01 to .87 for mathematics, and from .20 to .91 for science.

Table IV-8. Summary Statistics for Classical Item Difficulties for English Language Arts

Grade	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
3	47	.53	.11	.32	.46	.54	.59	.77
4	50 [a]	.53	.12	.27	.46	.52	.61	.78
5	46 [b]	.50	.09	.31	.45	.52	.56	.71
6	47	.52	.13	.16	.41	.51	.62	.78
7	47	.52	.12	.31	.43	.50	.59	.81
8	47	.52	.12	.17	.42	.53	.60	.77
10	46 [c]	.50	.09	.27	.43	.50	.58	.70

Note. P_{25} = 25th percentiles; P_{75} = 75th percentiles. [a] Grade-4 ELA has two operational forms: one for students who need accommodation and one for the general population. These two forms have only three item differences. [b] One grade-5 ELA item was removed from operational scoring. [c] One grade-10 ELA item was removed from operational scoring.

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	55	.50	.21	.10	.32	.49	.68	.87
4	55	.46	.16	.06	.35	.45	.56	.80
5	55	.47	.16	.10	.35	.48	.57	.80
6	55	.44	.17	.08	.32	.46	.57	.79
7	55	.45	.16	.01	.34	.48	.56	.72
8	55	.42	.20	.03	.25	.45	.56	.82
10	56	.40	.14	.09	.33	.40	.49	.70

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

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	35	.53	.15	.20	.42	.49	.61	.91
8	40	.45	.10	.24	.38	.44	.53	.67
11	40	.48	.12	.23	.40	.49	.58	.70

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

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 lower raw 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. The information on measuring the full performance continuum is not directly provided by classical test theory (CTT) item discrimination, but a test with more high-discrimination items will provide more-accurate measures of proficiency than a test with lower discriminating items.

Table IV-11, Table IV-12, and Table IV-13 present item discrimination for the three subjects. The means of item discrimination across grades range from .38 to .44 for ELA, from .39 to .48 for mathematics, and from .35 to .41 for science. For all subjects and grades, the minimums of item discrimination are over .15 except grade 4 in ELA, which is .14.

Table IV-11. Summary Statistics for Classical Item Discrimination for English Language Arts

Grade	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
3	47	.44	.08	.27	.39	.44	.49	.61
4	50 [a]	.39	.11	.14	.33	.39	.47	.63
5	46 [b]	.40	.11	.23	.31	.38	.48	.65
6	47	.38	.12	.13	.31	.38	.43	.62
7	47	.38	.08	.21	.31	.38	.43	.59
8	47	.40	.10	.20	.32	.39	.46	.61
10	46 [c]	.38	.10	.22	.31	.37	.47	.58

Note. P_{25} = 25th percentile; P_{75} = 75th percentile. [a] Grade-4 ELA has two operational forms: one for students who need accommodation and one for the general population. These two forms have only three item differences. [b] One grade-5 ELA item was removed from operational scoring. [c] One grade-10 ELA item was removed from operational scoring.

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	55	.48	.09	.27	.43	.47	.54	.63
4	55	.48	.09	.31	.42	.47	.53	.70
5	55	.46	.07	.35	.41	.45	.51	.63
6	55	.47	.09	.24	.40	.48	.53	.66
7	55	.44	.10	.23	.38	.43	.50	.67
8	55	.41	.09	.18	.35	.42	.45	.60
10	56	.39	.10	.16	.32	.39	.45	.61

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	35	.41	.08	.28	.34	.42	.46	.59
8	40	.35	.08	.19	.30	.36	.42	.50
11	40	.40	.09	.16	.34	.40	.47	.60

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

IV.3.2. Item Response Theory Item Statistics

KAP uses the two-parameter logistic IRT model and its polytomous counterpart, the graded response model, as measurement models. For those two IRT models, item parameters include item difficulty (i.e., *b* parameter) and item discrimination (i.e., *a* parameter). Section III.3.2. Item Response Theory and Model Assumptions has more-detailed information about these two IRT models and their item parameters.

Table IV-14, Table IV-15, and Table IV-16 summarize the difficulty (i.e., *b* parameter) estimates of operational items in ELA, mathematics, and science tests, respectively. IRT *b* parameter ranges from negative infinity to positive infinity and is on the same scale as the ability estimates. The higher the *b* parameter, the more difficult the item. 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 Table IV-14, Table IV-15, and Table IV-16.

The grade, mean IRT item difficulties range from −0.61 to −0.32 for ELA, from −0.41 to 0.23 for mathematics, and from −0.33 to 0.04 for science. For all grades and subjects, the ranges of item difficulties are large, ranging from −6.78 to 4.88 for ELA, from −7.32 to 3.33 for mathematics, and from −5.10 to 3.41 for science. The large IRT item-difficulty ranges indicate that the items included in KAP assessments adequately cover the full performance continuum.

Table IV-14. Summary Statistics for Item Response Theory Item Difficulty for English Language Arts

Grade	No. of *b* parameters	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
3	55	−0.51	0.85	−2.81	−0.99	−0.52	−0.12	1.71
4	61	−0.61	1.03	−4.02	−1.12	−0.65	−0.01	1.77
5	56	−0.36	1.43	−6.21	−0.96	−0.23	0.09	4.88
6	55	−0.43	1.15	−3.34	−1.20	−0.58	0.22	3.58
7	54	−0.32	1.22	−3.84	−1.00	−0.48	0.15	4.43
8	55	−0.42	1.47	−6.78	−1.05	−0.56	0.28	3.82
10	52	−0.32	0.95	−3.43	−0.84	−0.35	0.25	2.62

Note. *b* = difficulty parameter; P_{25} = 25th percentile; P_{75} = 75th percentile.

Table IV-15. Summary Statistics for Item Response Theory Item Difficulty for Mathematics

Grade	No. of *b* parameters	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
3	63	−0.41	1.40	−3.88	−1.38	−0.29	0.58	2.16
4	69	−0.08	1.20	−2.98	−0.80	−0.03	0.76	3.07
5	60	−0.21	1.04	−3.27	−0.70	−0.13	0.35	1.73
6	67	−0.24	1.52	−4.97	−0.87	−0.18	0.71	2.89
7	64	−0.08	1.13	−2.90	−0.74	−0.15	0.62	3.33
8	73	0.14	1.65	−3.99	−0.81	0.00	1.26	3.19
10	67	0.23	1.67	−7.32	−0.21	0.35	1.27	3.18

Note. *b* = difficulty parameter; P_{25} = 25th percentile; P_{75} = 75th percentile.

Table IV-16. Summary Statistics for Item Response Theory Item Difficulty for Science

Grade	No. of *b* parameters	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
5	40	−0.33	1.47	−5.1	−0.73	−0.13	0.50	3.41
8	40	0.04	0.74	−1.12	−0.55	−0.08	0.52	1.63
11	44	0.00	1.18	−3.82	−0.50	−0.17	0.49	3.12

Note. *b* = difficulty parameter; P_{25} = 25th percentile; P_{75} = 75th percentile.

Table IV-17, Table IV-18, and Table IV-19 summarize the discrimination (i.e., *a* parameter) estimates of a items in ELA, mathematics, and science tests, respectively. The IRT *a* parameter reflects an item's ability to differentiate students of high ability from those of low ability. Higher values indicate better discrimination power. As with CTT item discrimination, the information measuring the full performance continuum is not directly provided by IRT *a* parameters, but a test with more items having high item discrimination will provide more-accurate measures of proficiency than a test with fewer discriminating items.

The means of IRT item discrimination range from 0.80 to 1.07 for ELA, from 1.01 to 1.19 for mathematics, and from 1.77 to 0.89 for science. For all subjects and grades, the minimums of item discrimination are over 0.30, except for grade-6 ELA and grade-10 mathematics, which are 0.24 and 0.28, respectively. Although item discrimination is not usually too far from 1.0 on

average, the parameter clearly varies over items, justifying the use of the 2PL model, which permits the discrimination parameter to vary over items. Overall, mathematics has better discrimination parameters than ELA and science.

Table IV-17. Summary Statistics for Item Response Theory Item Discrimination for English Language Arts

Grade	No. of *a* parameters	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
3	49	1.07	0.33	0.54	0.82	0.98	1.31	1.86
4	52	0.92	0.38	0.33	0.60	0.87	1.22	1.94
5	46	0.86	0.36	0.30	0.58	0.80	1.07	2.01
6	47	0.86	0.37	0.24	0.60	0.72	1.05	1.79
7	47	0.80	0.31	0.42	0.56	0.71	0.98	1.73
8	47	0.91	0.33	0.35	0.68	0.89	1.18	1.74
10	46	0.80	0.28	0.34	0.61	0.77	0.98	1.61

Note. *a* = discrimination parameter; P_{25} = 25th percentile; P_{75} = 75th percentile.

Table IV-18. Summary Statistics for Item Response Theory Item Discrimination for Mathematics

Grade	No. of *a* parameters	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
3	55	1.19	0.33	0.56	0.91	1.19	1.38	1.87
4	55	1.16	0.33	0.70	0.87	1.11	1.30	2.02
5	55	1.19	0.32	0.56	0.95	1.20	1.40	1.93
6	55	1.18	0.34	0.60	0.88	1.13	1.46	1.91
7	56	1.11	0.37	0.45	0.85	1.01	1.31	1.96
8	55	1.04	0.32	0.33	0.83	1.01	1.26	1.76
10	56	1.01	0.40	0.28	0.71	0.97	1.28	2.00

Note. *a* = discrimination parameter; P_{25} = 25th percentile; P_{75} = 75th percentile.

Table IV-19. Summary Statistics for Item Response Theory Item Discrimination for Science

Grade	No. of *a* parameters	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
5	35	0.89	0.29	0.51	0.66	0.81	1.12	1.51
8	40	0.77	0.23	0.33	0.59	0.75	0.92	1.21
11	40	0.86	0.33	0.31	0.66	0.87	0.96	1.82

Note. *a* = discrimination parameter; P_{25} = 25th percentile; P_{75} = 75th percentile.

IV.3.3. Cognitive Complexity

KAP assessment items are categorized by cognitive complexity, as described by Webb's DOK model (Webb, 1997).

- Level 1 (recall) requires simple recall of information, such as a fact, definition, term, or simple procedure.
- Level 2 (skill/concept) involves some mental skills, concepts, or processing beyond a habitual response. Students must make some decisions about how to approach a problem or activity. Keywords distinguishing a level 2 item include "classify," "organize," "estimate," "collect data," and "compare data."

- Level 3 (strategic thinking) requires reasoning, planning, using evidence, and thinking at a higher level.
- Level 4 (extended thinking) requires complex reasoning, planning, developing, and thinking, most likely over an extended time. Cognitive demands are high, and students are required to make connections both within and among subject domains.

The DOK associated with each cluster 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.

For the new 2022 grade-10 math assessment, item cognitive complexity is considered an interaction between the item and the student's abilities. When applying DOK to assessment items, we considered the cognitive path that test takers engage in when responding to test items. The cognitive path to respond to an item for each student can vary according to several factors, including, but not limited to, engagement with the item, prior experiences, and solution paths (Ackerman, 1987; Anderson, 1992, 1996; Wine & Hoffman, 2022). These variations can lead to different levels of cognitive complexity for different students when responding to a single item because we must consider how all test takers engage with the item. For some test takers, proficiency in knowledge, skills, or abilities may lower the cognitive complexity of the task. Conversely, instability in the knowledge, skills, and abilities can prompt cognition at a higher DOK level (Logan, 1985; Wine & Hoffman, 2022). Thus, the possible DOK ranges for an item were:

- Level 1–2 (recall or skill/concept): Recitation or recognition of facts, basic reading comprehension, use of algorithms or procedures. Includes recitation or identification of explanations learned previously. Automatic or rote application of a skill warrants DOK level 1. Some degree of inference and analysis, basic decision-making, performance of work without strategic planning, selection of the correct simple tool or procedure and its application. Conscious and deliberate decision-making warrants DOK level 2.
- Level 2–3 (skill/concept or strategic thinking): Some degree of inference and analysis, basic decision-making, performance of work without strategic planning, selection of the correct simple tool or procedure and its application. Conscious application of the skills warrants DOK level 2, which is the explanation of decisions, thinking process, or work performed; strategic planning or the application of multipart reasoning to determine a course of action; and citing evidence to support reasoning. Deliberate strategizing of how to combine skills warrants DOK level 3.

Table IV-20 shows the percentage of operational items by DOK level or range for each 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 grades 3–8 mathematics, most items are at level 1 and level 2 as well, with relatively fewer level 3 items. All grade-10 mathematics items are at level 1, level 2, or level 1–2. For science, most items are at levels 2 and 3, with a few items at level 1 and level 4.

Table IV-20. Percentage of Items by Depth of Knowledge (DOK) Level, Subject, and Grade

Grade	English language arts (ELA)				Mathematics					Science				
	DOK level, %				DOK level, %					DOK level, %				
	Total items	1	2	3	Total items	1	2	3	12	Total items	1	2	3	4
3	47	26	60	15	55	62	38	0						
4	50 [a]	22	70	8	55	51	47	2						
5	46 [b]	26	59	15	55	64	36	0		35	0	71	29	0
6	47	30	47	23	55	56	44	0						
7	47	11	83	6	55	60	38	2						
8	47	23	72	4	55	51	44	5		40	5	43	53	0
10	46 [c]	2	91	7	56	41	29	0	30					
11										40	3	50	45	3

Note. [a] Grade-4 ELA has two operational forms: one for students who need accommodation and one for the general population. These two forms have only three item differences. [b] One grade-5 ELA item was removed from operational scoring. [c] One grade-10 ELA item was removed from operational scoring.

IV.4. Scoring and Scaling

This section introduces the procedures of scoring individual items, scoring the test as a whole, and scaling. We include test results and the performance-level distribution for 2022 KAP testing and present the KAP performance trend for the previous five years. Finally, this section describes the quality-control procedures used to ensure the accuracy of scoring and scaling.

IV.4.1. Scoring

Item and test scoring in the 2022 administration remained the same as in previous years. All items were machine scored. The same test-scoring method used previously was used this year.

IV.4.1.1. Item Scoring

All KAP assessment items administered in 2022 were machine scored. The online test-delivery platform compared student responses to the correct keys stored with the items and assigned the scores accordingly.

IV.4.1.2. Test Scoring

Test scoring used a psychometric model to derive item scores on the test to produce a single score indicating a student's proficiency level. We computed the IRT ability estimates (i.e., thetas) using the 2PL model and GRM. Because the total score was derived using the summed-score method (Thissen & Wainer, 2001)—in which scores for each item were added together to derive the raw score—thetas had a one-to-one correspondence with raw scores (i.e., each raw score has only one matching theta). By using the test-characteristic curve function of the IRT models, we obtained the theta for each raw-score point for a test form (Press et al., 1988).

IV.4.2. Scaling

Scaling is the process of transforming thetas or raw scores to a reporting scale. The purpose of scaling is to facilitate the use and interpretation of test scores. The scale is also the basis for reporting 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. However, it can be difficult to use and interpret negative test scores with decimal points. To support score interpretation, it is useful to transform thetas to a scale composed of positive integers. The next section addresses the process for constructing scale scores.

IV.4.2.1. Scale Transformation

Kolen and Brennan (2004) used the following formula to derive scaling constants:

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

where $SS(y)$ is the scale 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 scale-score equivalent to the original score, y_1. This equation can be structured as

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

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

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

A and C are the slope and intercept of the scaling constants, respectively. The Kansas State Department of Education (KSDE) predetermined the scale score to have a slope, A, of 25 for all subjects and grades.

The KAP assessment has four performance levels: 1, 2, 3, and 4. Higher performance levels indicate higher performance on the test. Students in levels 3 or 4 are considered to have met the academic expectation of postsecondary readiness. KSDE determined a scale score of 300 to be the cut that separates Levels 2 and 3, and standard-setting panels set the original theta values of Level 2 and level 3 cuts of each subject and grade. With the original cut score (y_1), equivalent scale score (i.e., $SS[y_1] = 300$), and a scale-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.4.2.2. Scale-Transformation Constant

The test-scoring process described in Section IV.4.1.2. estimates the IRT thetas for students. Then, the y_1 in Equation IV-6 is the theta associated with a scale score of 300. Standard-setting panels set the grade theta cuts for ELA and mathematics (grade 3–8) in 2015, theta cuts for science in 2017, and theta cuts for grade-10 mathematics in 2022 (see theta cuts in Table IV-21, Table IV-22, and Table IV-23).

We find the C for each grade using Equation IV-6. Table IV-24 shows the scale-transformation constants for all grades and subjects. Because A and C are known, we can derive the other two scale-score cuts using Equation IV-4. The derived scale-score cuts may have decimal points. The final operational scale cut scores need to be rounded to a possible integer scale score, depending

on the rounding rule (Cizek et al., 2004). Note that for ELA and mathematics, except grade-10 mathematics, the rounding rule is that the scale-score cuts are rounded up. 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. For grade-10 mathematics and science, the rounding rule is that the scale-score cuts are rounded to the nearest integer because grade-10 mathematics and science already have very rigorous standards and the scale-score cuts should not be increased through rounding up.

Table IV-21. English Language Arts Cut Scores

Grade	Theta cuts			Scale-score cuts		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 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			Scale-score cuts		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 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	−0.420	0.660	1.830	273	300	329

Table IV-23. Science Cut Scores

Grade	Theta cuts			Scale-score cuts		
	Level 2	Level 3	Level 4	Level 2	Level 3	Level 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. English Language Arts (ELA), Mathematics, and Science Scaling Constants

Grade	ELA		Mathematics		Science	
	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	283.50		
11					25	292.13

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

IV.4.2.3. Properties of Scale scores

The derived scale 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. Software typically assigns those raw-score points a theta of −99 or 99. To keep the scale score meaningful, the lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS) are set to cap scale scores within a reasonable range. KAP's LOSS and HOSS are set as 220 and 380, respectively.

IV.4.3. Operational Test Results

This section presents the results of the 2022 administration of the KAP, including descriptive statistics representing the number of students tested by various subgroups; the 2022 scale-score summary for all students and by subgroup; the 2022 performance-level distribution for each subject by grade; and the 2022 participation data, scale-score summary, and proficiency rates compared to those of previous years. This report includes participation rates prominently because it is critical to account for variability in participation when interpreting KAP performance within and across years.

IV.4.3.1. Student Participation

In 2022, states administered the KAP operational test in ELA, mathematics, and science in grades 3–8 and high school. At the high school level, students completed ELA and mathematics assessments in grade 10 and science assessments in grade 11. As described in Section I.3. Required Assessments and Intended Population, Kansas is committed to including all students in the KAP assessment.

Table IV-26 shows the number of enrolled students and of tested students, as well as participation rate by subject and grade. The definitions for the indicators are:

- *Enrolled students* are students assigned to take a KAP test.
- *Tested students* are students receiving a score report. Students receive a score report when they were not exempt (exemption rules are described in Section I.3. Required Assessments and Intended Population), complete at least five items in each of the two test

sections, and have logged out of the testing platform for the first section. This reporting rule has been used since 2015.

- The *participation rate* is calculated as the number of tested students divided by the number of enrolled students.

As shown in Table IV-25, more than 33,000 students were tested for each subject and grade. Across all subjects and grades, the participation rates ranged from 96% to 99%. The participation rate in elementary and middle school grades was greater than 98%, especially at elementary grades (about 99%). High school grades had a lower participation rate, with 97% for ELA, 97% for mathematics, and 96% for science. Across all subjects and grades, the average participation rate was 98%.

Table IV-25. Number and Participation Rate (PR) of Enrolled and Tested Students by Subject and Grade

Grade	English language arts			Mathematics			Science		
	Enrolled (*N*)	Tested (*N*)	PR (%)	Enrolled (*N*)	Tested (*N*)	PR (%)	Enrolled (*N*)	Tested (*N*)	PR (%)
3	35,356	35,016	99%	35,389	35,068	99%	-	-	-
4	35,878	35,524	99%	35,907	35,557	99%	-	-	-
5	35,799	35,461	99%	35,830	35,480	99%	35,849	35,540	99%
6	36,953	36,470	99%	36,968	36,438	99%	-	-	-
7	37,371	36,799	98%	37,388	36,783	98%	-	-	-
8	38,173	37,492	98%	38,191	37,478	98%	38,204	37,547	98%
10	36,747	35,659	97%	36,799	35,584	97%	-	-	-
11	-	-	-	-	-	-	35,259	33,908	96%

Table IV-26 shows participation rates by student group [11] and by School Board of Education (SBOE) district. The participation rates by student group and by SBOE district are not subject specific. If a student participated in one subject of the KAP assessment, then the student is included in the calculation. The 286 school districts in Kansas are distributed among 10 SBOE districts. Some school districts appear in multiple SBOE districts when district boundaries reach into more than one SBOE district. The Kansas Unified School Districts document lists the school districts included in each SBOE district. Comparing the participation rates of students within each subject and grade by gender, ethnicity, race, EL status, and disability status, we note the following results:

- no difference in participation rates between gender groups
- very similar participation rates for different race groups except in high schools
  - Black and NHPI students have lower participation rates than Asian and White students in high schools
- a slightly higher participation rate for non-Hispanic students than for Hispanic students in high schools
- a slightly higher participation rate for ELs than for non-ELs in elementary schools
- a slightly higher participation rate for students without disabilities than for students with disabilities, especially in high schools
  - Students without disability have a 4% higher participation rate than students with disability in high schools.

The comparison of participation rates of different SBOE districts within each grade showed the following results:

- Participation rates in elementary schools are very similar across districts.
- District 4 has slightly lower participation rates in middle schools than other districts.
- Districts 1, 4, 8, and 10 have slightly lower participation rates in high schools.

Districts 1 and 4 include the Kansas City, Topeka, and Lawrence school districts. District 8 includes the Wichita school district. Appendix C provides detailed demographic distribution of SBOE districts.

Table IV-26. Participation Rate by Demographic Characteristics and State Board of Education (SBOE) District

Characteristic	Grade							
	3 (%)	4 (%)	5 (%)	6 (%)	7 (%)	8 (%)	10 (%)	11 (%)
Gender								
Female	99	99	99	98	98	97	96	96
Male	99	99	99	98	98	98	96	96
Race								
Native American	99	99	100	98	98	97	95	96
Asian	97	97	98	98	97	98	98	97
Black	98	98	98	97	95	95	93	93
NHPI	96	100	98	98	98	96	94	92
Other	99	98	99	98	98	96	94	95
White	99	99	99	98	98	98	97	97
Hispanic								
No	99	99	99	98	98	98	97	96
Yes	99	99	99	98	98	97	95	95
Student with disability								
No	99	99	99	99	98	98	97	97
Yes	98	98	98	97	97	96	93	93
English learner								
No	99	99	99	98	98	98	96	96
Yes	98	98	99	97	97	97	94	95

SBOE district								
1	99	98	98	96	97	96	94	94
2	99	99	99	99	98	98	98	97
3	99	99	98	99	98	98	97	97
4	98	98	98	95	96	95	93	93
5	99	99	99	99	99	99	98	98
6	99	100	99	99	99	99	98	98
7	99	99	99	98	98	97	96	96
8	98	98	98	97	97	96	94	94
9	99	99	99	99	99	98	97	97
10	99	99	99	98	98	97	95	95

Note. NHPI = Native Hawaiian and Pacific Islander.

For all tested students, Table IV-27 shows the percentage of students in each student group by grade. This summary is not subject specific. If a student tested in one subject of the KAP assessment, then the student is included in the calculation. The student groups include gender, race, ethnicity, disability status, and EL status. [4] The percentages of students in each student group were very similar across grades except students with disability. There were approximately equal percentages of male and female students. The largest percentage tested by race group was White, and the largest percentage tested by ethnicity group was non-Hispanic. More students without disability were tested than students with disability, and more non-ELs were tested than ELs. There was a decrease in percentage of students with disability across grades. The lower grades had higher percentages of students with disability than did higher grades.

[4] Economically disadvantaged status is not shared with ATLAS to protect the privacy of students, so this student group is not included in the comparison.

Table IV-27. Percentage of Tested Students by Demographic Characteristic and Grade

Characteristic	Grade							
	3 (%)	4 (%)	5 (%)	6 (%)	7 (%)	8 (%)	10 (%)	11 (%)
Gender								
Female	49.00	48.89	49.14	48.88	48.89	49.11	48.97	49.53
Male	51.00	51.11	50.86	51.12	51.11	50.89	51.04	50.47
Race								
Native American	1.97	1.90	1.99	2.13	2.32	2.63	2.94	3.29
Asian	3.01	3.00	2.98	2.91	2.93	2.88	2.96	3.04
Black	6.98	6.90	6.96	7.12	7.35	7.39	6.99	6.91
NHPI	0.33	0.36	0.33	0.34	0.30	0.26	0.26	0.26
Other	7.80	7.35	7.37	7.01	7.04	6.89	6.60	6.63
White	79.91	80.49	80.37	80.49	80.06	79.96	80.26	79.88
Hispanic								
No	78.98	79.10	79.02	78.99	78.45	78.43	79.07	79.70
Yes	21.02	20.90	20.98	21.01	21.55	21.57	20.94	20.30
SWD								
No	80.64	81.45	82.08	83.27	84.54	85.39	86.63	87.62
Yes	19.36	18.55	17.92	16.73	15.46	14.61	13.37	12.38
EL								
No	86.60	86.60	87.82	89.18	89.78	91.66	92.43	92.37
Yes	13.40	13.40	12.18	10.82	10.22	8.34	7.57	7.63

Note. NHPI = Native Hawaiian and Pacific Islander; SWD = student with disability; EL = English learner.

IV.4.3.2. Operational Test Results

Table IV-28, Tables IV-29, and Table IV-30 present summaries of scale scores by grade for ELA, mathematics, and science. As noted previously, it is critical to consider variability in participation rates when interpreting KAP performance within and across years.

The minimum and maximum scale scores for each grade and subject were 220 and 380, respectively. As shown in Tables IV-28 through IV-30, the mean scale scores were close to 300 in lower grades (i.e., grades 3–6 in ELA, grades 3–4 in mathematics, and grade 5 in science) and approximately 280 in higher grades. The standard deviations of scale scores were very similar across grades within one subject. Science tends to have higher standard deviations of scale scores than ELA and mathematics.

Table IV-28. Scale-Score Descriptive Statistics for English Language Arts

Grade	*M*	*SD*	Min.	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max.
3	292.7	27.5	220	260	270	290	311	331	380
4	296.7	27.8	220	263	276	294	316	334	380
5	293.5	29.3	220	259	272	290	314	332	380
6	288.3	28.6	220	250	266	288	308	327	380
7	286.1	29.1	220	248	264	283	307	326	380
8	277.9	27.4	220	244	256	276	295	313	380
10	279.9	29.4	220	242	258	278	301	320	380

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

Table IV-29. Scale-Score Descriptive Statistics for Mathematics

Grade	*M*	*SD*	Min.	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max.
3	300.0	30.8	220	260	277	298	321	341	380
4	291.2	29.1	220	257	268	288	309	332	380
5	289.3	28.3	220	257	268	283	307	328	380
6	286.3	28.5	220	254	264	282	305	327	380
7	286.2	27.5	220	256	267	280	303	326	380
8	281.3	27.0	220	253	262	276	296	318	380
10	282.1	26.7	220	255	264	276	294	320	380

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

Table IV-30. Scale-Score Descriptive Statistics for Science

Grade	*M*	*SD*	Min.	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max.
5	299.0	31.5	220	263	276	295	321	343	380
8	281.8	28.2	220	251	263	278	299	319	380
11	287.6	30.3	220	254	266	282	306	329	380

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

Table IV-31 and Figure IV-4, Figure IV-5, and Figure IV-6 provide the percentage of students achieving each performance level (i.e., level 1 through level 4) and the proficiency rate (i.e., percentage at level 3 and level 4) by subject and grade. Proficiency rates across all subjects and grades ranged from 21% to 49%. All three subjects tended to have lower proficiency rates in higher grades. A summary of the results across grades by subject follows.

- ELA
  - Level 1 percentages ranged from 19% to 38%.
  - Level 2 percentages ranged from 29% to 43%.
  - Level 3 percentages ranged from 19% to 33%.
  - Level 4 percentages ranged from 3% to 14%.
  - As grades increased, level 1 and level 2 percentages tended to increase and level 3 and level 4 percentages tended to decrease.

- Mathematics
  - Level 1 percentages ranged from 20% to 48%.
  - Level 2 percentages ranged from 28% to 49%.
  - Level 3 percentages ranged from 14% to 31%;.
  - Level 4 percentages ranged from 4% to 19%.
  - As grades increased, level 1 percentages tended to increase and level 3 and level 4 percentages tended to decrease.
  - Level 2 percentages tended to be stable across grades.
- Science
  - Level 1 percentages ranged from 27% to 43%.
  - Level 2 percentages ranged from 28% to 29%.
  - Level 3 percentages ranged from 16% to 27%.
  - Level 4 percentages ranged from 8% to 18%.
  - As grades increased, level 1 percentages tended to increase and level 3 and level 4 percentages tended to decrease.
  - Level 2 percentages tended to be stable across grades.

Table IV-31. Percentage of Students Achieving at Each Performance Level (PL) for English Language Arts (ELA), Mathematics, and Science

Grade	ELA PL (%)					Mathematics PL (%)					Science PL (%)				
	1	2	3	4	P	1	2	3	4	P	1	2	3	4	P
3	32	30	25	13	38	23	28	31	19	49					
4	19	38	33	10	43	20	45	25	10	36					
5	31	29	26	14	40	34	35	20	11	31	27	29	27	18	44
6	37	29	27	6	33	36	35	21	9	30					
7	37	32	25	7	32	24	49	23	4	27					
8	36	43	19	3	21	48	31	17	4	21	47	29	16	7	24
10	38	36	21	5	26	45	34	14	7	21					
11											43	28	18	11	29

Note. P = proficiency (combination of performance levels 3 and 4). Column percentages may not total 100 because of rounding.

Figure IV-4. Performance-Level Distribution for English Language Arts

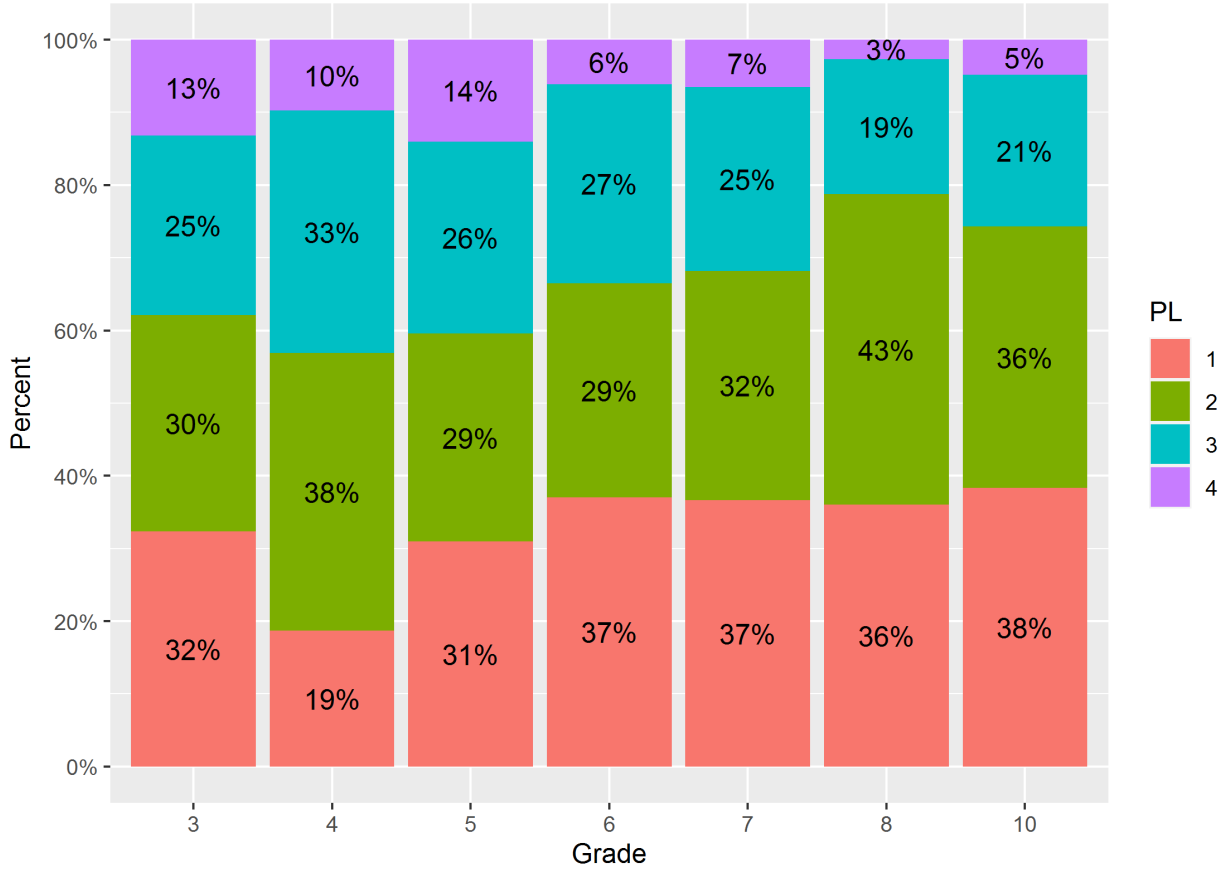


Figure IV-5. Performance-Level Distribution for Mathematics

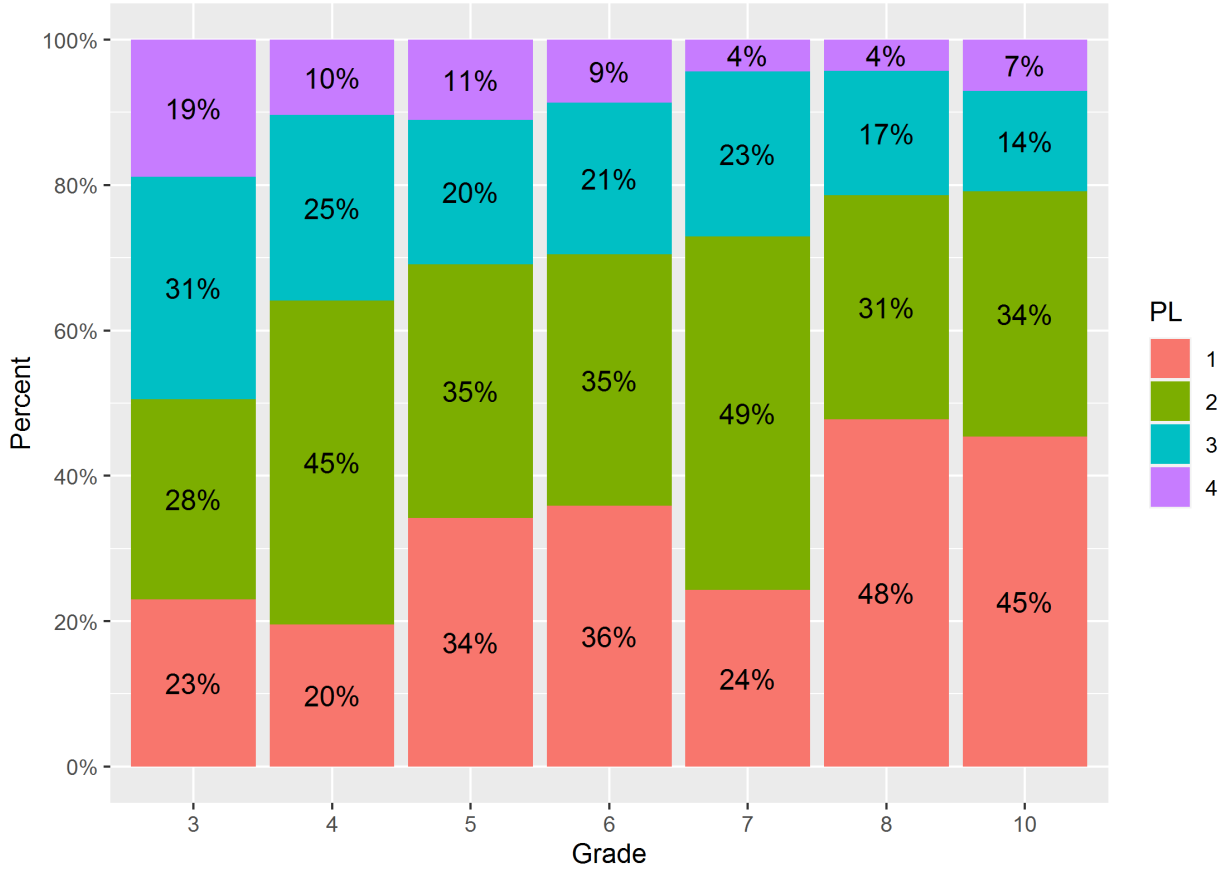


Figure IV-6. Performance-Level Distribution for Science

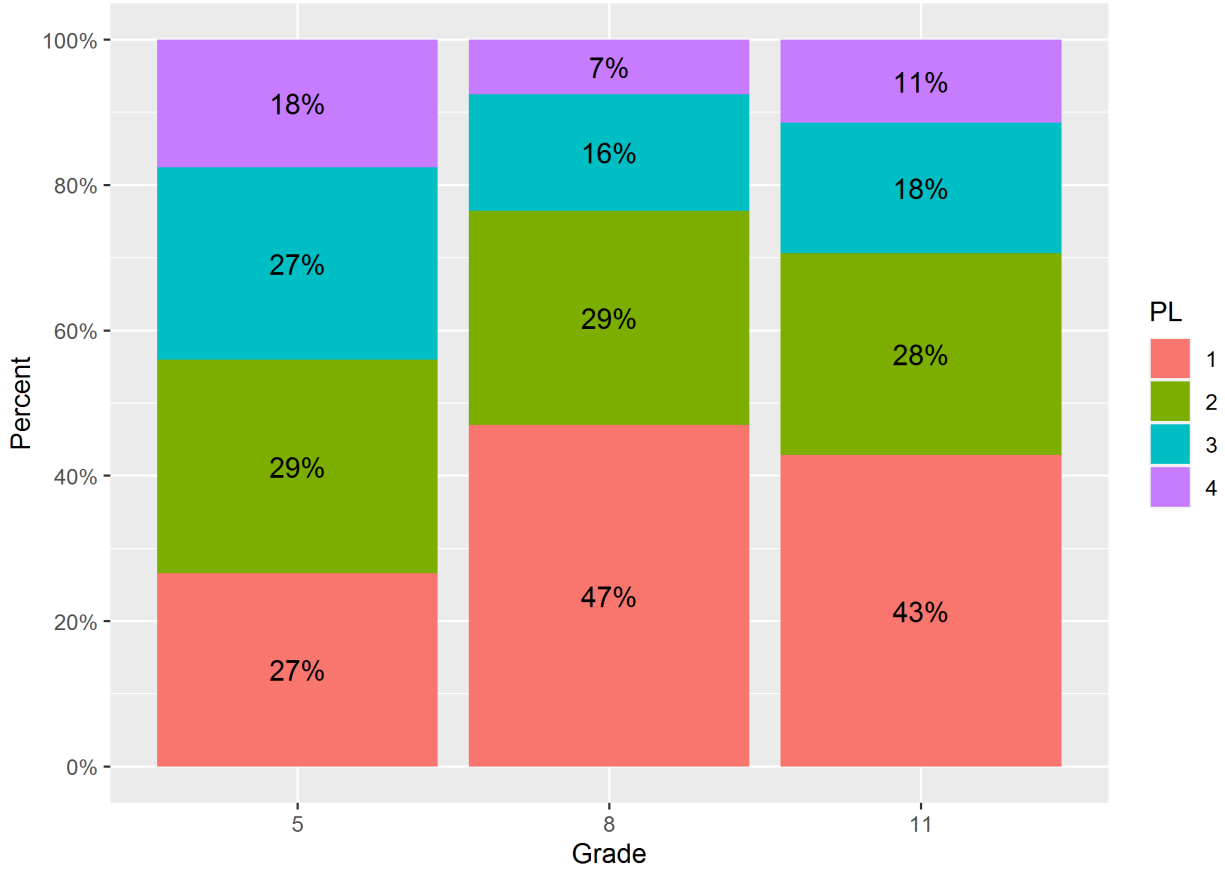


Table IV-32, Table IV-33, and Table IV-34 summarize the mean and standard deviation of the scale scores by demographic student group.[5] For all subjects and grades, the mean scale score was above 280 and the standard deviation was around 30. The comparison of scale-score mean and the standard deviation of different student groups within each subject and grade indicate that female students scored higher in ELA and male students scored slightly higher in mathematics and science. Male students had higher standard deviations, Asian students had the highest means and standard deviations, and Black students had the lowest means and standard deviations. Non-Hispanic students had higher means and standard deviations than Hispanic students, non-ELs had higher means and standard deviations than ELs, and students without disabilities had higher means and standard deviations than students with disabilities.

[5] Economically disadvantaged status is not shared with ATLAS to protect the privacy of students, so this student group is not included in the comparison.

Table IV-32. English Language Arts Mean and Standard Deviation of Scale Scores by Grade and Student Subgroup

Subgroup	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 10	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Gender														
Male	291.2	27.1	295.1	27.7	291.6	29.0	285.5	29.0	283.0	29.1	274.2	26.7	275.8	29.1
Female	294.1	27.7	298.4	27.8	295.4	29.4	291.2	27.8	289.3	28.8	281.7	27.7	284.1	29.0
Race														
NA	278.8	22.1	284.9	24.0	282.9	24.6	277.3	25.5	273.9	25.6	266.4	23.3	268.4	25.2
Asian	299.4	28.9	306.6	29.8	304.5	32.6	299.1	30.3	298.9	31.0	289.6	30.3	291.2	31.6
Black	278.7	23.0	281.3	24.5	277.8	24.8	272.4	25.5	271.6	25.4	264.0	23.3	264.4	25.0
NHPI	282.4	23.7	288.9	26.3	282.2	27.6	280.7	27.2	271.4	23.8	271.4	26.4	274.3	23.9
Other	288.8	26.3	293.5	26.6	290.7	29.2	284.2	28.3	283.0	28.4	275.5	26.7	277.0	28.6
White	294.4	27.5	298.3	27.7	295.0	29.1	289.9	28.3	287.6	29.0	279.3	27.3	281.4	29.3
Hispanic														
Yes	282.3	23.2	285.8	24.9	282.0	25.1	277.4	25.7	275.3	25.9	268.1	24.1	269.8	26.4
No	295.4	27.8	299.6	27.9	296.5	29.5	291.1	28.6	289.0	29.2	280.6	27.7	282.5	29.5
SWD														
Yes	275.9	22.7	278.8	24.2	273.8	25.2	266.0	24.6	263.6	24.4	256.2	21.2	255.9	21.9
No	296.6	27.0	300.7	27.0	297.7	28.3	292.7	27.2	290.1	28.0	281.5	26.7	283.4	28.7
EL														
Yes	278.0	21.1	281.6	23.2	275.3	22.0	268.6	22.0	264.9	21.4	256.5	18.5	254.2	18.6
No	294.8	27.6	299.0	27.8	296.0	29.2	290.6	28.4	288.4	28.9	279.8	27.3	281.9	29.1

Note. NA = Native American; NHPI = Native Hawaiian and Pacific Islander; SWD = student with disability; EL = English learner.

Table IV-33. Mathematics Mean and Standard Deviation of Scale Scores by Grade and Student Subgroup

Subgroup	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 10	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Gender														
Male	302.4	32.3	294.1	30.4	292.0	30.2	287.0	29.4	288.2	29.0	282.1	28.0	282.2	27.8
Female	297.5	28.9	288.3	27.4	286.6	25.8	285.5	27.4	284.1	25.8	280.5	25.9	281.9	25.5
Race														
NA	285.2	27.1	279.5	25.3	280.0	23.0	277.2	24.2	275.0	20.1	270.5	21.3	271.5	18.4
Asian	312.0	33.4	304.4	33.5	306.3	34.1	302.4	35.7	304.5	35.6	299.3	35.4	301.5	38.3
Black	281.1	26.5	272.6	22.3	272.8	20.5	269.3	22.2	271.2	19.4	267.5	20.5	268.4	18.1
NHPI	290.0	29.3	281.8	26.4	280.9	21.3	277.9	23.3	272.4	20.4	275.8	24.9	277.5	26.8
Other	294.9	29.9	286.1	26.6	285.3	27.8	280.6	26.8	281.0	25.0	277.8	25.6	278.0	26.0
White	302.1	30.4	293.1	29.0	290.7	28.1	287.9	28.2	287.7	27.5	282.6	26.8	283.2	26.4
Hispanic														
Yes	287.8	26.8	279.2	23.9	279.1	22.6	275.0	23.7	275.3	21.5	271.4	21.7	272.7	20.6
No	303.2	30.9	294.4	29.5	292.0	29.0	289.3	28.9	289.2	28.3	284.0	27.6	284.5	27.6
SWD														
Yes	281.1	29.2	274.6	25.0	273.0	23.1	266.2	22.6	267.2	20.0	262.9	19.5	264.4	17.4
No	304.4	29.4	295.0	28.6	292.8	28.0	290.2	27.8	289.6	27.3	284.4	26.8	284.7	26.9
EL														
Yes	284.6	26.2	277.0	24.0	275.5	20.7	268.4	20.4	268.8	17.5	264.5	17.9	265.0	14.7
No	302.3	30.7	293.4	29.2	291.3	28.6	288.4	28.5	288.1	27.8	282.8	27.1	283.4	27.0

Note. NA = Native American;; NHPI = Native Hawaiian and Pacific Islander; SWD = student with disability; EL = English learner.

Table IV-34. Science Mean and Standard Deviation of Scale Scores by Grade and Student Group

Subgroup	Grade 5		Grade 8		Grade 10	
	M	*SD*	*M*	*SD*	*M*	*SD*
Gender						
Male	300.6	33.1	284.1	29.6	288.6	32.4
Female	297.3	29.6	279.4	26.5	286.6	27.8
Race						
Native American	288.0	27.2	269.1	23.7	274.3	24.4
Asian	310.2	34.5	290.5	30.1	298.9	34.8
Black	281.1	25.7	266.1	21.4	270.6	22.5
NHPI	284.5	29.0	271.8	26.2	275.4	23.5
Others	295.9	31.4	279.2	26.8	284.5	29.6
White	300.7	31.3	283.5	28.3	289.4	30.3
Hispanic						
Yes	287.4	27.1	270.7	24.0	275.4	25.0
No	302.0	31.9	284.8	28.5	290.6	30.7
Student with disability						
Yes	281.8	28.9	264.7	23.8	268.0	23.7
No	302.7	30.8	284.6	27.9	290.3	30.1
English learner						
Yes	281.4	25.1	261.3	18.6	263.8	17.4
No	301.4	31.5	283.6	28.2	289.5	30.3

Note. NHPI = Native Hawaiian and Pacific Islander.

IV.4.3.3. Participation Trend

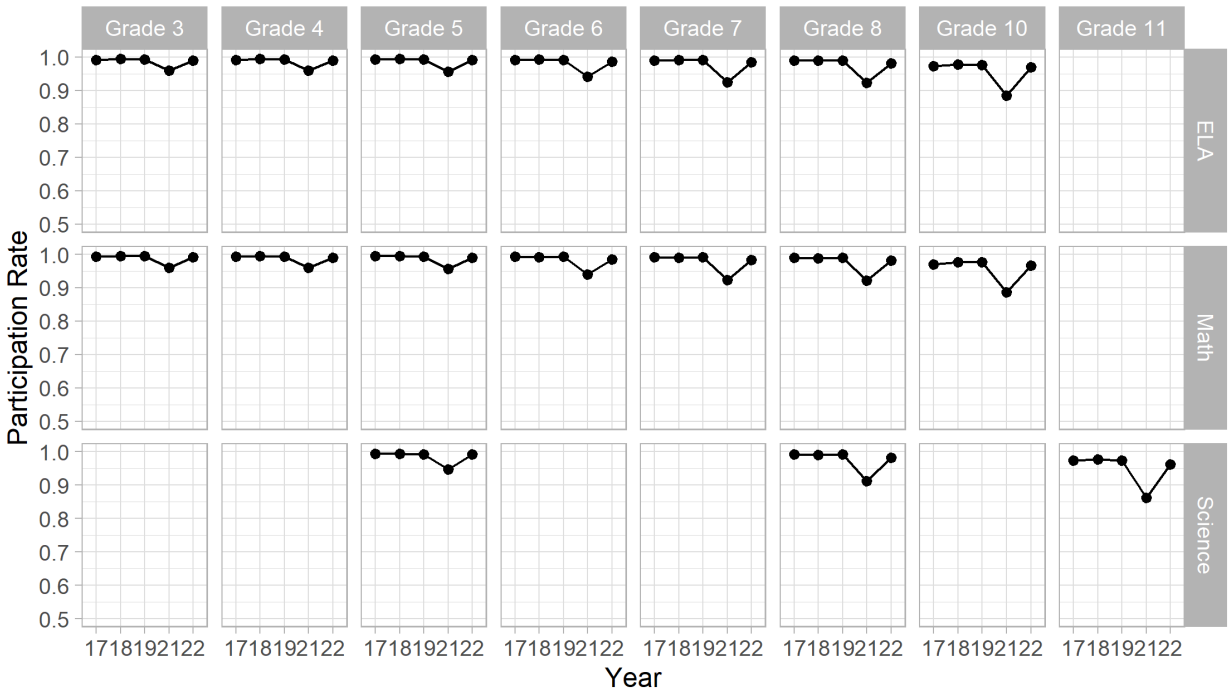
Table IV-35 presents enrollment trends for 2017–2022 for ELA, mathematics, and science. The numbers were very similar in the higher grades across years; however, in grades 3, 4, and 5, there was a decrease of approximately 3,000 enrolled students from 2019–2021 per subject and grade; the number of enrolled students became stable from 2021–2022, except in grade 5 with a decrease of 1,000 enrolled students from 2021 to 2022. When comparing the enrollment numbers in a student cohort (i.e., enrollments in grade 3 in 2017, grade 4 in 2018, grade 5 in 2019, grade 7 in 2021, grade 8 in 2022), the enrollment numbers were very stable, with a slight decrease (fewer than 700 students) in 2021.

Table IV-35. Total Number of Enrolled Students by Subject and Grade for 2017–2022

Subject	Grade	2017	2018	2019	2021	2022
English language arts	3	38,599	37,724	37,316	35,440	35,356
	4	38,707	38,600	37,920	35,547	35,878
	5	37,761	38,532	38,606	36,735	35,799
	6	37,098	37,655	38,537	37,225	36,953
	7	37,132	37,018	37,680	38,145	37,370
	8	36,990	37,114	37,065	38,275	38,173
	10	36,382	36,245	36,973	36,811	36,747
Mathematics	3	38,612	37,792	37,346	35,455	35,389
	4	38,704	38,653	37,950	35,557	35,907
	5	37,773	38,576	38,619	36,743	35,830
	6	37,120	37,704	38,561	37,224	36,968
	7	37,141	37,064	37,693	38,142	37,387
	8	37,010	37,179	37,076	38,286	38,191
	10	36,395	36,292	36,994	36,813	36,799
Science	5	37,785	38,615	38,632	36,756	35,849
	8	37,026	37,203	37,103	38,301	38,204
	11	34,929	34,976	34,938	35,527	35,259

Figure IV-7 presents the participation rates (i.e., proportion of students receiving a score report out of students enrolled) for different subjects and grades by year from 2017–2022. From 2017–2019, the participation rates were approximately 98% for all grades. There was a decrease in participation rates from 2019 to 2021, from approximately 98% to 93% in lower grades and from approximately 98% to 88% in higher grades. Then in 2022, the participation rates increased to 98% for all grades compared to 2021.

Figure IV-7. Participation Rates for 2017–2022 by Subject and Grade



IV.4.3.4. Performance Trend

ELA, mathematics, and science mean scale-score trends for 2017–2022 are presented in Figure IV-8. Note that grade 10 mathematics mean scale score for 2022 is not included because it is a new assessment with a new IRT scale. For ELA in grades 3, 4, 5, and 6, the mean scale scores decreased from 2017 to 2018, increased in 2019, and decreased in 2021 and 2022. For ELA in grades 7, 9, and 10, there was a slight decrease in mean scale scores from 2017 to 2022. The average difference in mean scale score was approximately 1 scale-score point across grades between years. For mathematics, the mean scale scores decreased from 2017 to 2018, increased from 2018 to 2019, decreased from 2019 to 2021, and increased from 2021 to 2022. For grade-5 science, the mean scale increased slightly from 2017 to 2018, decreased slightly from 2018 to 2021, and increased from 2021 to 2022. For grade-8 science, the mean scale scores decreased from 2017 to 2022. For grade-11 science, the mean scale scores decreased from 2017 to 2019, increased from 2019 to 2021, and decreased from 2021 to 2022.

Figure IV-8. Longitudinal Scale-Score Trend by Subject and Grade for 2017–2022

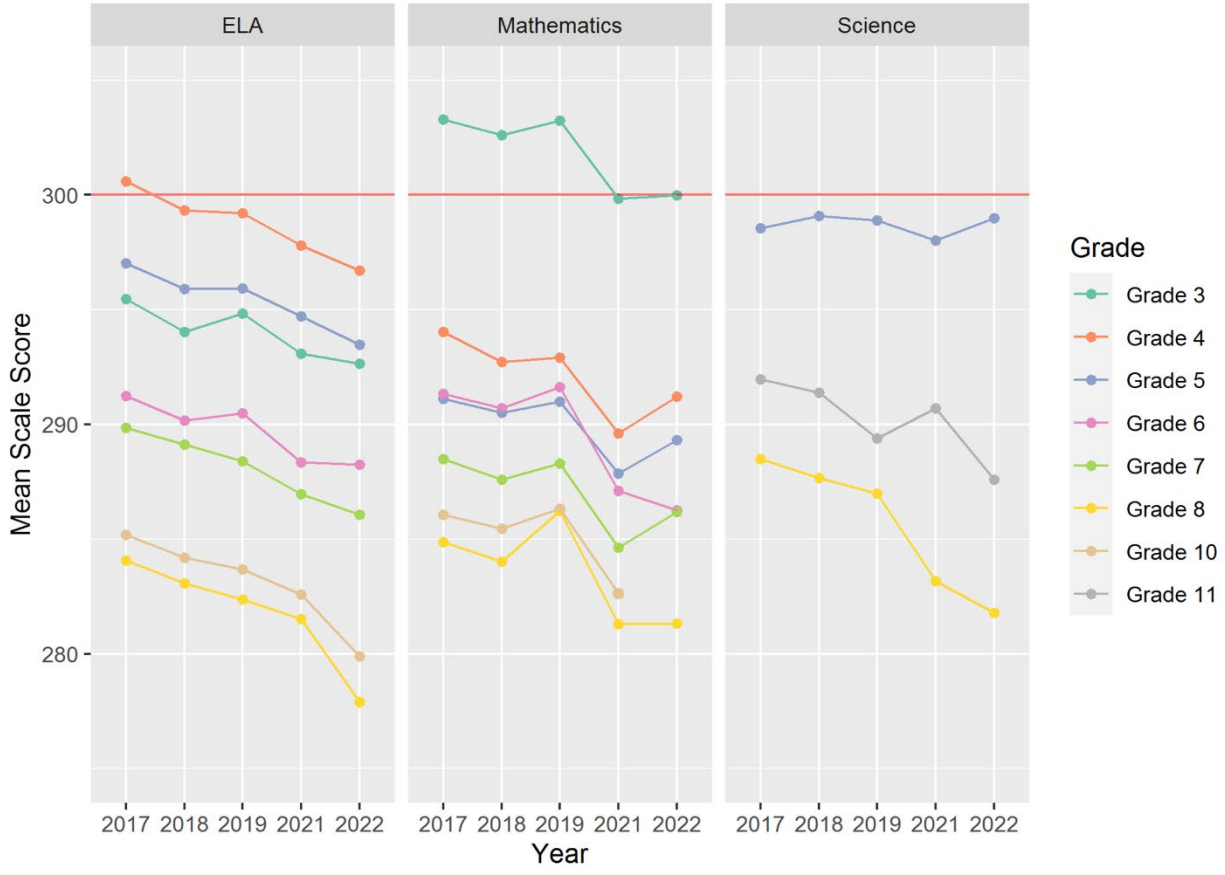
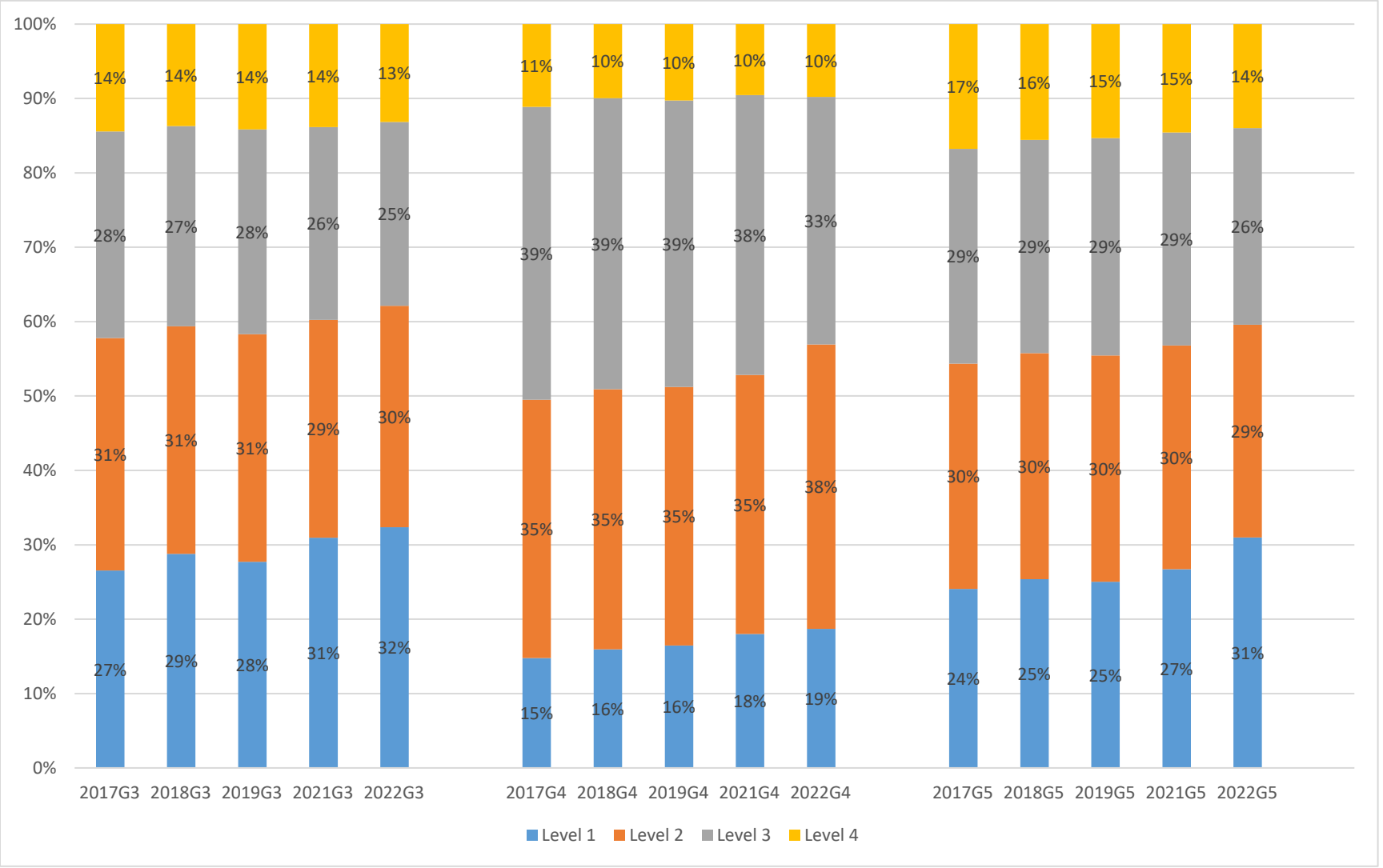


Figure IV-9, Figure IV-10, Figure IV-11, Figure IV-12, and Figure IV-13 present the performance-level distribution trends across years for ELA, mathematics, and science. Table IV-36, Table IV-37, and Table IV-38 present the proficiency-rate trends across years for ELA, mathematics, and science. Grade-10 mathematics performance-level distribution for 2022 is not included because it is a new assessment with a new IRT scale and cut scores. (Details about setting the new cut scores for grade-10 mathematics are in Section VI.2.2.2. 2022 Grade-10 Mathematics Standard Setting.) A summary of the results across grades by subject follows.

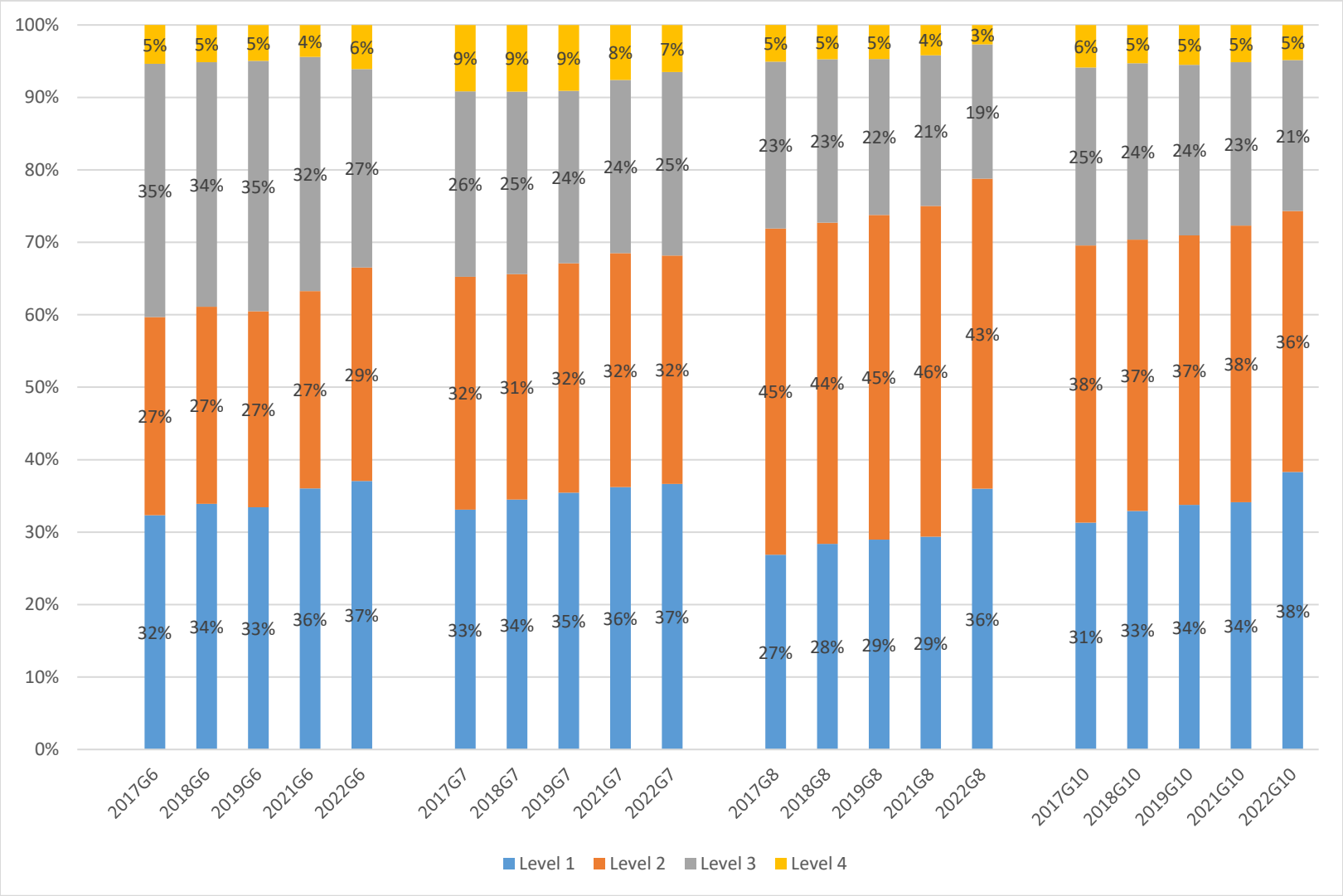
- ELA
  - There was an increase in percentage of level 1 students,
  - a very stable percentage of level 2 students,
  - a decrease in percentage of level 3 students,
  - and a stable percentage of level 4 students from 2017–2022.
  - For the proficiency rate, there was a slight decrease in proficiency rates from 2017–2022.
- Mathematics
  - There was an increase in percentage of level 1 students except in 2019,
  - a very stable percentage of level 2 students,
  - a decrease in percentage of level 3 students especially in 2021 and 2022,
  - and a stable percentage of level 4 students, with an increase in level 4 percentage in 2019.
  - Proficiency rates in most elementary grades decreased from 2017–2021 and increased in 2022.
  - Proficiency rates in middle school and high school grades increased in 2019, decreased in 2021, and leveled in 2022.
- Science
  - The grade-5 performance-level distributions were similar between 2017 and 2018 and between 2019, 2021, and 2022.
  - The grade-5 proficiency rates were slightly lower in 2019, 2021, and 2022 compared with previous years.
  - Grade-5 science had a larger level 4 percentage in 2022 than previous years.
  - For grade-8 science, there was a decrease in proficiency rates from 2017–2022.
  - For grade-11 science, 2017, 2018, and 2021 had very similar performance-level distributions, and 2019 and 2022 had a decrease in proficiency rates.

Figure IV-9. Performance-Distribution Trend for English Language Arts for Grades 3–5



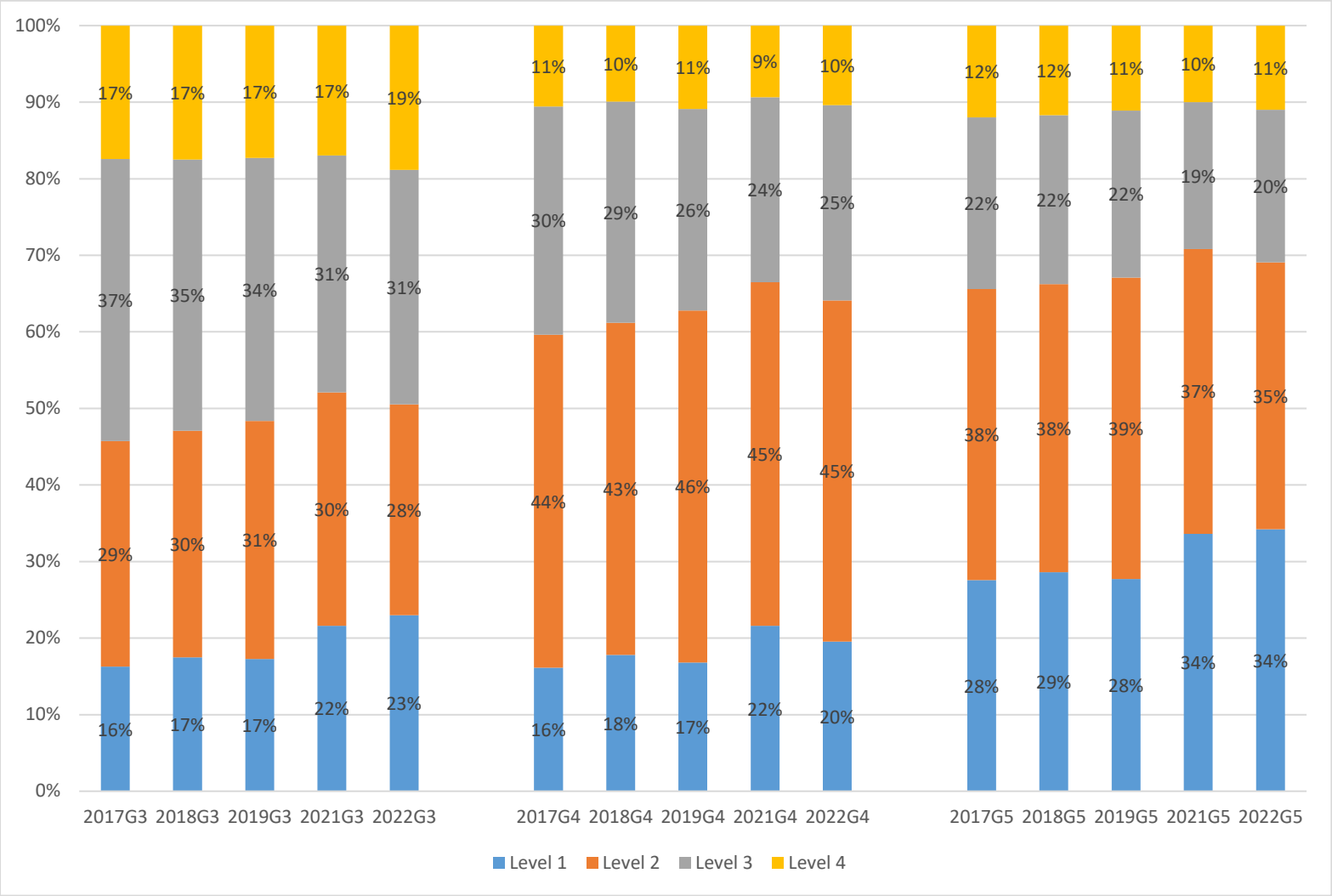
Note. G = grade.

Figure IV-10. Performance-Level Distribution Trend for English Language Arts for Grades 6–10



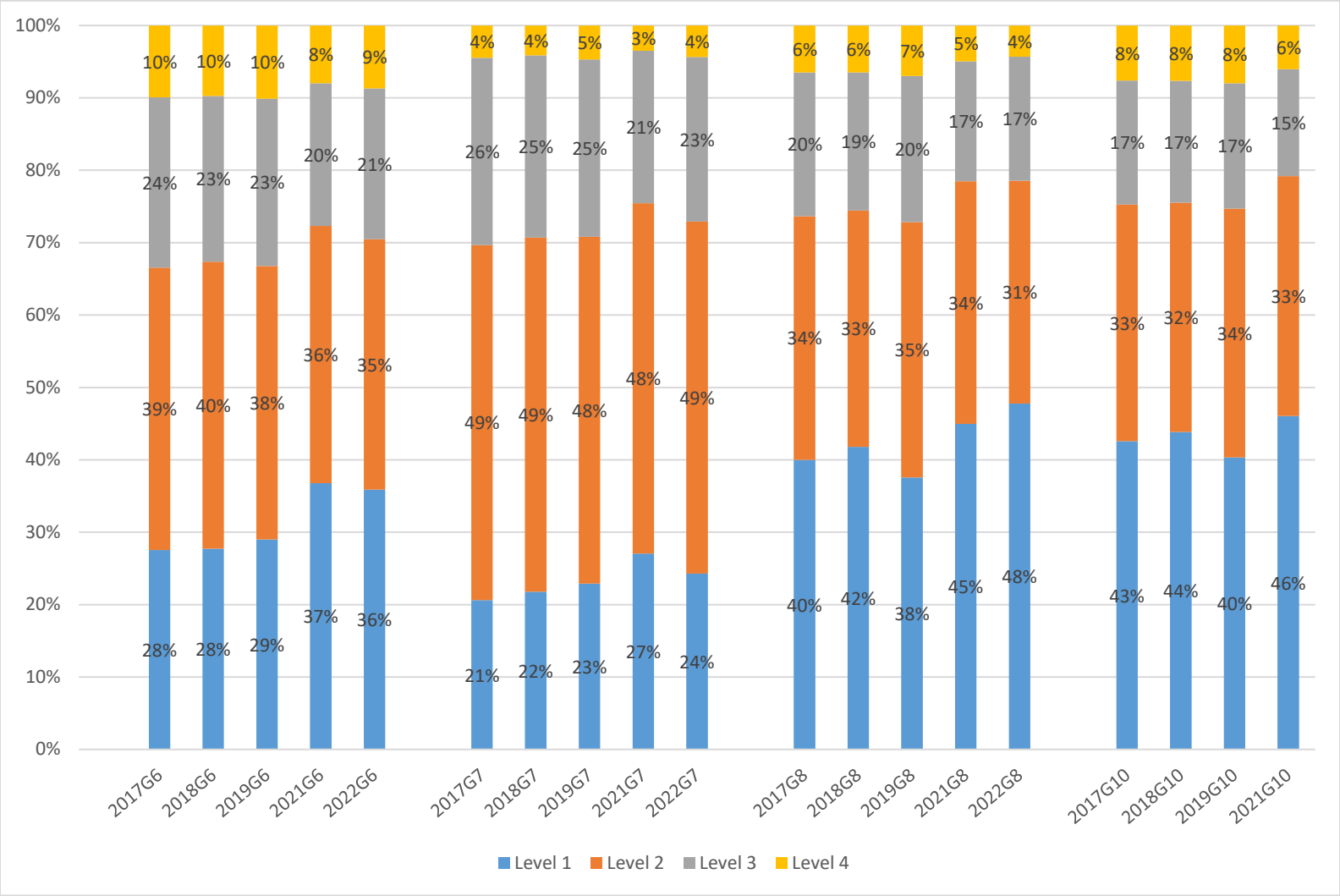
Note. G = grade.

Figure IV-11. Performance-Level Distribution Trend for Mathematics for Grades 3–5



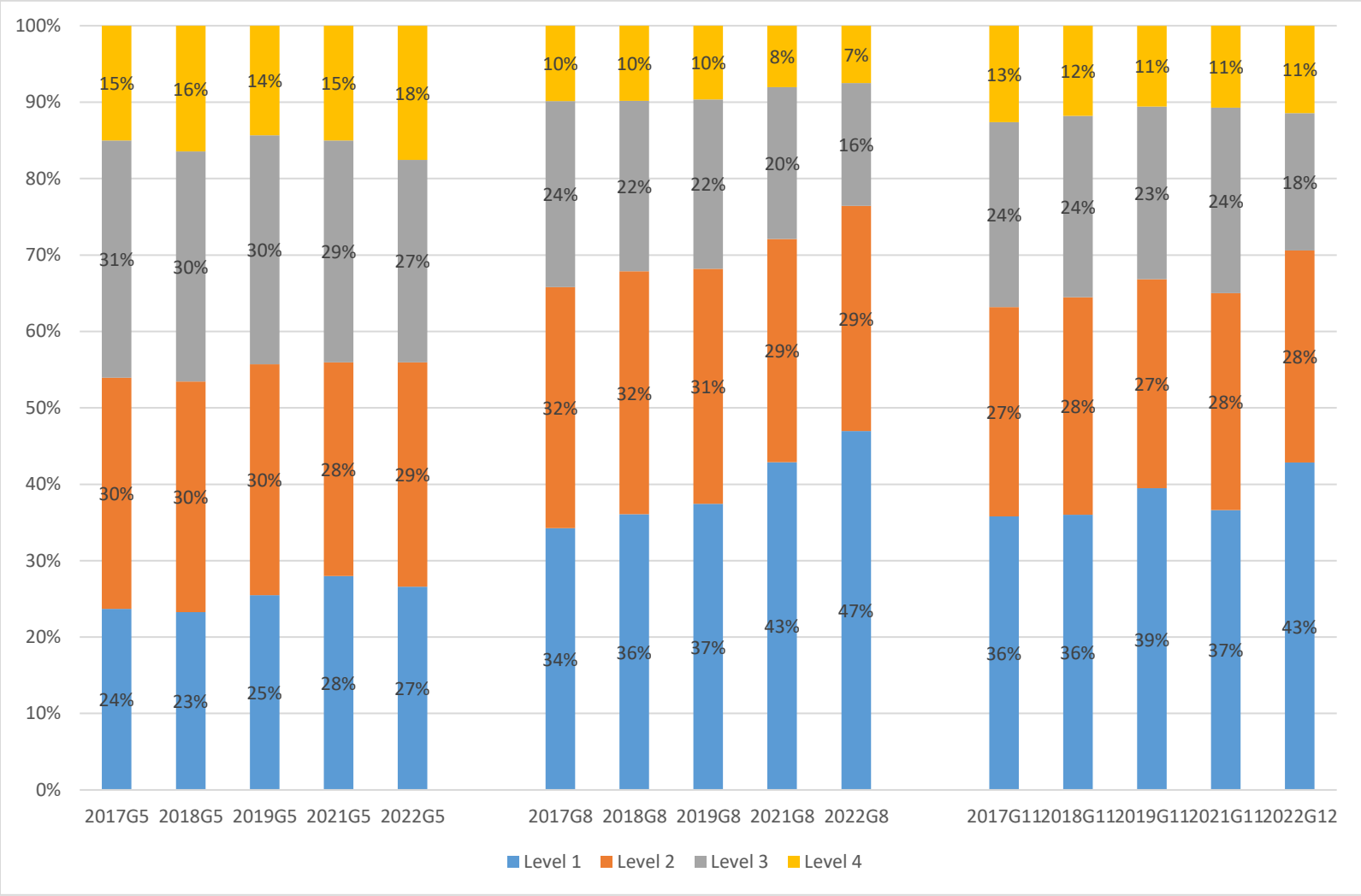
Note. G = grade.

Figure IV-12. Performance-Level Distribution Trend for Mathematics for Grades 6–10



Note. G = grade.

Figure IV-13. Performance-Level Distribution Trend for Science



Note. G = grade. Column percentages may not total 100% because of rounding.

Table IV-36. Proficiency Rates for English Language Arts, 2017–2022

Year	Grade %						
	3	4	5	6	7	8	10
2017	42	50	46	40	35	28	30
2018	41	49	44	39	34	27	30
2019	42	49	45	40	33	26	29
2021	40	47	43	37	32	25	28
2022	38	43	40	33	32	21	26

Table IV-37. Proficiency Rates for Mathematics, 2017–2022

Year	Grade %						
	3	4	5	6	7	8	10
2017	54	40	34	33	30	26	25
2018	53	39	34	33	29	26	24
2019	52	37	33	33	29	27	25
2021	48	34	29	28	25	22	21
2022	49	36	31	30	27	21	--

Table IV-38. Proficiency Rates for Science, 2017–2022

Year	Grade %		
	5	8	11
2017	46	34	37
2018	47	32	36
2019	44	32	33
2021	44	28	35
2022	44	24	29

IV.4.3.4.1. Monitoring the COVID-19 Effect

In 2022, we continued to monitor the effects of COVID-19 on classroom instruction as reported by teachers in the annual teacher survey. Among the 277 educators (approximately 1% of educators in Kansas) who responded to the instruction questions on the teacher survey, 223 (79%) agreed or somewhat agreed that their students received instruction that was similar to a typical year before the COVID-19 pandemic, 54 (19%) disagreed or somewhat disagreed that their students received typical instruction, and five (2%) said that this question was not applicable to them. Educators who responded that their students had not received typical instruction also described the main differences in instruction and learning experiences. The differences included spending instructional time to fill knowledge gaps; students missing instruction because of quarantine, mask mandates and social distancing affecting learning experiences; and changes in students' attitudes and attention toward learning. These survey results suggest that COVID-19 may still affect instruction and learning experiences for some students in Kansas.

IV.4.3.5. Quality-Control Checks

The scoring and reporting process of KAP test results had multiple quality-control steps. First, student-response data were checked at least three times during the testing window for scoring errors or duplicates.

Second, we conducted classical item analysis during the testing window using approximately 20% of the overall test volume. We compared the calculated classical item statistics from the current year's data with classical item statistics obtained from data of previous years. The purpose of this step was to monitor the classical item statistics trend and ensure items were functioning as expected. During 2022 classical item analysis, one grade-5 ELA item did not function as expected because of a change in how the item was presented, which differed from its presentation in field testing. In consultation with KSDE, we removed this item from scoring. One grade-7 mathematics item did not function as expected because the calculator status was different from when the item was field tested. We treated this math item as an operational field-test item and updated its item parameters using data obtained in the 2022 administration.

Third, we recalibrated IRT statistics for items and calculated classical statistics using this year's data after the window closed. We compared both newly calculated IRT and classical item statistics with statistics from years when the items were field tested. This analysis and comparison allowed us to evaluate item drift. In 2022, no items were flagged as unstable using item-stability flagging criteria established for the KAP program. However, in reviewing the classical item-difficulty statistic plots, we found that the item-difficulty changes of some ELA items (three to 10 items per grade out of 47 items) were larger than other ELA items. These items had formats updated to be more accessible this year (content remained identical, however; e.g., drag-and-drop items were converted to matrix items to allow for switch use). Other ELA items did not have any format changes because they were already widely accessible. Given the change in formatting for these items, we treated the updated ELA items as operational field-test items and updated their item parameters according to the 2022 administration. This recalibration process ensures that the formatting change did not negatively affect student test scores.

Fourth, two psychometric staff members independently generated and compared scoring tables. They examined reasonableness and accuracy of the scoring tables through predetermined criteria:

- All subjects and grades were represented.
- All tests were represented.
- All raw scores were represented for each form.
- No integer was missing from the scale scores, from 0 to the maximum test form score.
- The scale score increased with the raw score within each form.
- The minimum scale score was 220, and the maximum scale score was 380.

Fifth, at least two psychometric staff members independently checked the cut scores used to classify students to ensure they were consistent with the cut scores approved by the Kansas State Board of Education.

Sixth, we calculated and compared the summary statistics of testing results with those of previous years to ensure the performance trend was reasonable.

Finally, the psychometric and technology teams independently calculated each individual student's total score, scale score, performance levels, subscores, and subscore performance levels. We compared results from the two teams' independent calculation to identify any differences or calculation errors. We generated students' score reports only after the scoring results from both teams were identical. The purpose of all quality-control steps was to ensure the scoring results provided on students' reports were complete and accurate.

IV.5. Multiple Assessment Forms

In large-scale assessment programs, different item sets may be used on test forms 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 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.

All three subject areas used one operational form in 2022, so their linking involves cross-year equating procedures. In grade-10 mathematics, 2022 is the first operational administration for the new assessment aligned to the 2017 Kansas Standards. Only one operational form was developed and administered; thus, no linking was conducted for grade-10 mathematics in 2022.

IV.5.1. Cross-Year Linking Design

To increase the number of linking items and maximize linking stability, the cross-year linking uses the preequating method. All items on the 2022 ELA, mathematics (except grade 10), and science tests have IRT parameters calibrated, and they are on the same IRT scale as items in 2015 for ELA and mathematics (except grade 10) and 2017 for science. 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 on the base scale (i.e., the 2015 scale for ELA and mathematics [except grade 10] and the 2017 scale for science).

IV.5.2. Cross-Year Linking Procedure

All items on the 2022 ELA, mathematics (except grade 10), and science tests were field tested in previous years. In those years, all field-tested items were calibrated using concurrent item calibration after the test window closed by fixing the item parameters of the operational items so that the field-test item parameters were placed on the same IRT scale as operational items, that is, base scale. Also, the test characteristics such as TIF and scoring tables are compared across years during test-construction psychometric review to ensure that different test forms across years have similar test characteristics, that is, similar reliability estimates and similar raw-score cuts.

IV.6. Multiple Versions of an Assessment

The KAP is administered online via the Kite platform, which can be used on PCs with Windows, Macs, Chromebooks, and iPads. All students who take the KAP must use the Kite Student Portal (described in Section II.4.2. Test-Administration Procedures). The Kite platform can provide various accommodations for students with special needs. For details about available accommodations, please refer to Section V.4. Accommodations. The one exception is that a paper–pencil braille form is provided to students who need it. No grade or subject-area test has

more than 10 students taking the braille form [6]. The braille version has the same operational items as the online version but no field-tested items. When the American Printing House (APH) translated items to braille format, it modified some formats of items to provide adequate experience for students who are blind or visually impaired without introducing construct-irrelevant variance. For example, the radio buttons of the selected-response items on the online version are changed to option labels (e.g. A, B, C, and D). Moreover, APH and the AAI content team collaborate to construct the test-administration notes for the braille form, which add clarifying language so that students who are blind or visually impaired can access the same information as their sighted peers.

IV.7. Technical Analysis and Ongoing Maintenance

This technical manual includes a series of technical analyses that use this year's testing data. These analyses include DIF analysis, relationships among different assessment, reliability analyses, classification consistency and accuracy analyses, test result summary, and trend analysis.

In addition to the technical analyses, this technical manual also contains the analysis for ongoing monitoring of the effect of COVID-19. We collected and summarized the contextual data about instruction from the KAP teacher survey in this technical manual in Section IV.4.3.3.1. Monitoring the COVID-19 Effect. Survey results indicated that COVID-19 still affects instruction of some students in Kansas.

Student-response data were checked at least three times during the testing window for scoring errors or duplicates to ensure all items were scored correctly and data were captured correctly. Moreover, classical item analysis of all items were conducted when there are 20% of the overall test volume to make sure items were functioning as expected. During classical item analysis, option analysis was also conducted for multiple-choice (keyed or multiple selection) items. In 2022–2023, we plan to expand the item-analysis process to include checking specific features or characteristics of technology-enhanced items (e.g., response rates of each part for multipart items). We will review every type of technology-enhanced item on the operational KAP assessment and identify features or characteristics in those item types that need to be monitored during item analysis.

[6] The sample sizes of braille forms were too small to undertake a comparability study between the braille version and online version.

V. Inclusion of All Students

This chapter presents information about the inclusion of all students in the Kansas Assessment Program (KAP), including students with disabilities and English learners (ELs). The procedures for including students with disabilities and ELs are summarized, followed by a description of the available accessibility tools and accommodations. More information about accessibility supports and accommodations for KAP can be found in the *Kansas Accessibility Manual*, Tools and Accommodations for the Kansas Assessment Program, and the *Kansas Assessment Examiner's Manual 2021–2022*.

The Kansas State Department of Education (KSDE) complies with the Elementary and Secondary Education Act (ESEA) and the Individuals with Disabilities Education Improvement Act (IDEA), both of which require all students, including students with disabilities and ELs, to participate in assessments used for accountability purposes. One of the principles of ESEA is strong 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), Section 504 plan, or individual learning plan (ILP). All of these plans are developed to address each student's unique needs.

V.1. Procedures for Including Students With Disabilities

Accessibility tools and accommodations that are available either within or outside the Kite® system allow students with disabilities to take KAP assessments. Details about different tools and accommodations are in Section V.3. Accessibility Tools and Section V.4. Accommodations. The inclusion of students with disabilities is achieved by providing clear guidelines for educators, so they can register their students with different needs. The *Kansas Assessment Examiner's Manual 2021–2022* *describes* step-by-step registration procedures for students who need accommodations.

V.2. Procedures for Including English Learners

As described in Section I.3. Required Assessments and Intended Population, ELs are required to take the KAP assessments, although they do not have to take the English language arts (ELA) test in the first year. Accessibility tools and accommodations that are available either within or outside the Kite system allow ELs to take KAP assessments. Specific accessibility tools and accommodations for ELs include directions read aloud by a synthetic voice, electronic translators and word-to-word translators (not for ELA passages), translation dictionaries, and Spanish keyword translation for mathematics and science assessments. Details about different tools and accommodations are in Section V.3. Accessibility Tools and Section V.4. Accommodations. The inclusion of ELs is achieved by providing clear guidelines for educators, so they can register their students with different needs. The *Kansas Assessment Examiner's Manual 2021–2022* describes step-by-step registration procedures for students who need accommodations.

V.3. Accessibility Tools

Accessibility tools are available for all students taking KAP assessments and vary by subject. Table V-1 describes these tools and recommendations for use.

Table V-1. KAP Accessibility Tools

Tool	Description
Calculator: basic or TI-108	Allows students to perform simple mathematical calculations. Depending on test settings, the basic calculator icon will display either the basic calculator or the TI-108 Emulator. This tool is available for mathematics grades 6–8, and 10, and science grades 5, 8, and 11. May not be available in mathematics sections that measure numbers and operations.
Calculator: TI graphing	Allows students to plot graphs, solve equations, and display several lines of calculations on the screen. Available for grade-10 mathematics. May not be available in mathematics sections that measure numbers and operations.
Calculator: TI scientific	Allows students to perform calculations in science, engineering, and mathematics. Available for mathematics grades 6–8 and science grades 8 and 11. May not be available in mathematics sections that measure numbers and operations.
Eraser	Removes highlighting and striker marks from the screen.
Guide line	When selected, follows the student's pointer and highlights the text of a reading passage, line by line. Differs for iPads, where the line remains stationary as the student scrolls through the passages.
Highlighter	Allows students to select text on the screen and highlight the selected text with a pink background.
Mark for review: question answered	When selected by test takers, changes the item-number indicator at the top of the screen to blue, with an accompanying flag graphic.
Mark for review: question unanswered	When selected by test takers, changes the item-number indicator at the top of the screen to red, with an accompanying flag graphic.
Notes	Presents a yellow rectangle on the screen where students can type notes about the test content.
Periodic table	Presents a standard periodic table. Students can select an individual element to view the atomic number, atomic mass, and full element name. Default view shows elements by their abbreviations. Tool is available for science tests.
Pointer	Allows students to select items in the test.
Search	Allows students to enter search terms; matching words are then highlighted in orange.
Striker	Allows students to place a line through an answer choice that is not desired.
Tags	Allows students to use various tags within a reading passage. Tags remain in the passage until the student selects Clear All. The

Tool	Description
	available tags are Main Idea, Supporting Details, Key Word, Evidence, Reread This, and Help.
TTS: directions	Allows students to have a synthetic voice read directions aloud on the assessment.
TTS: science	Allows students to have a synthetic voice read directions, stimuli, and test items aloud on the science assessment.
Whole-screen magnification	Allows students to magnify the screen up to four levels.
Sketch pad	Allows students to draw, write, create shapes, etc.

Note: TTS = text-to-speech audio.

V.4. Accommodations

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 designed to reduce or eliminate the effects of a student's disability or English proficiency; however, they do not alter learning expectations. The accommodations provided to a student, documented in a student's IEP, Section 504 plan, or ILP, should be the same for classroom instruction, classroom assessments, and local education 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.

It is important for educators to become familiar with state policies regarding accommodations during assessments. According to the *Kansas Assessment Examiner's Manual 2021–2022* (p. 23), reading to students any text (including isolated words) in the passages on the ELA test is prohibited. Only a very limited number of students, such as those who cannot access printed text, may be permitted to have passages read through text-to-speech (TTS) software with approval from KSDE staff. Another prohibited accommodation is for teachers and students to bring pregenerated journals and logs.

The *Kansas Assessment Examiner's Manual 2021–2022* provides more details regarding accommodations in KAP assessments, including an overview, prohibited practices, and recording accommodations used during testing (e.g., most testing accommodations should be entered into the student's Personal Needs Profile [PNP]). Additional information about accommodations or Kite toous are in t*he Kite Student Portal Manual for Test Administrators 2021–2022*. Table V-2 presents the accommodations available for KAP assessments.

Table V-2. Available Accommodations for KAP Assessments

Tool	Description
American Sign Language (ASL)	Displays available ASL videos for the assessment question; available for mathematics and science only.
Auditory calming	Provides relaxing, peaceful music that can play while the student takes the test.
Braille form [a]	Provides a paper–pencil braille test form.
Color contrast	Sets a text color and a background color. Options are gray text on black background, yellow text on black background, green text on white background, and red text on white background.
Color overlay	Provides a color background behind the content on the screen. Color options are light blue, light yellow, light gray, light red, and light green.
Key word translation	Provides Spanish translation for item keywords; available for mathematics and science only.
Masking (student controlled or presented by default)	Allows a student to mask, or cover, parts of the test. After a student selects the masking button, a black box appears. The student can move the masking box by dragging it to different areas of the screen.
Reverse contrast	Sets the text color to white and the background color to black.
Switches	Allows students to interact with the assessments through the use of a single switch or key instead of a mouse.
TTS: Items	Provides a synthetic voice that reads text and test items aloud.
TTS: Items and passages	Provides a synthetic voice that reads ELA passages aloud.
Whole-screen magnification	Allows students to magnify the screen according to what was set up in their Personal Needs Profile.

Note: ELA = English language arts; TTS = text-to-speech audio. [a] Starting in 2021–2022, a new braille online form fully aligns with the paper–pencil braille form.

V.4.1. Selection of 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 the KAP assessments. Second, students with IEPs, Section 504 plans, or ILPs 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 Kite Educator Portal before beginning any assessment. For TTS software requests, local test administrators need to enter the support in the Audio & Environment Support section of the student's PNP. The *Kite Educator Portal Manual for Test Coordinators* lists the steps for creating a PNP for students.

Test administrators handle some accommodations (e.g., braille, magnification device) that are allowed for the KAP assessment, but most accommodations (e.g., color contract) 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. For additional accommodations documented in a student's plan that are not available for KAP assessments, teachers should contact the District Test Coordinator (DTC), who will send the request to KSDE staff for approval. These additional requested accommodations should not change the construct being tested.

V.4.2. Frequency of Accommodation Use

The summary of PNP accommodation requests shown in Table V-3 indicates the number of students for whom each accommodation is requested. This table summarizes PNP selections by grade. Note that some students may receive multiple accommodations. The table shows that TTS: Items is the most commonly requested accommodation option. This accommodation makes an audio recording of the test item available.

Table V-3. Frequency of Accommodation Requests by Grade

Accommodation	Grade							
	3	4	5	6	7	8	10	11
American Sign Language (ASL)	14	11	19	24	11	20	10	14
Auditory calming	29	57	170	179	181	133	95	124
Braille form	1	3	0	2	4	5	5	7
Color contrast	6	8	14	16	15	14	5	21
Color overlay	5	7	25	27	23	20	19	28
Key word translation	108	237	271	303	349	321	333	298
Masking	8	7	13	10	8	11	4	10
Reverse contrast	3	0	4	3	2	6	4	6
Switches	1	3	2	10	2	1	11	7
TTS: Items	4,707	4,812	4,859	4,372	4,319	4,008	3,013	2,588
TTS: Items and passages	194	172	148	72	69	43	17	0
Whole-screen magnification	31	47	35	57	47	74	95	79
Total	5,113	5,372	5,574	5,091	5,045	4,670	3,616	3,203

Note: TTS = text-to-speech audio.

VI. Academic Achievement Standards and Reporting

This chapter describes updates related to achievement standards and reporting for the Kansas Assessment Program (KAP). For the subjects of English language arts (ELA) and mathematics (except for grade-10 mathematics), the KAP assessment uses the same achievement standards that were set in 2015; grade-10 mathematics uses new achievement standards that were set in 2022. For science, the assessment uses the same achievement standards that were set in 2017. The next sections describe the standard-setting procedure and outcomes for all subjects and grades. Different types of score reports and resources for the 2022 test administration are also described in this chapter.

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

Policy performance level descriptors (PLDs) define the KAP academic achievement standards. Although the KAP assessment is based on content standards, the assessment evaluates student performance using academic achievement standards. PLDs describe the expected academic achievement at each performance level. Classifying student assessment performance into a given performance level means that the student meets the minimum expected knowledge and skills of that performance level. This score interpretation applies to all students who participate in the KAP assessment. The policy PLDs have four levels: 1, 2, 3, and 4. Students who achieve Levels 3 and 4 are considered to have met the academic expectations of postsecondary readiness, that is, they met the proficiency. The state adopted the new academic achievement standards defined by the policy PLDs [7] for ELA and mathematics in grades 3–8 in 2015, for grade-10 mathematics in 2022, and for science in 2017.

VI.2. Achievement Standard Setting

For the KAP assessment, standard setting occurred in 2015 for ELA and mathematics and in 2017 for science. The 2022 KAP assessment continues to use the achievement standards that were set in 2015 for ELA and mathematics in grades 3–8 and in 2017 for science. However, for the 2022 grade-10 mathematics assessment, new achievement standards were established. The next sections describe the standard-setting method, procedure, and outcomes for all subjects and grades.

VI.2.1. Standard-Setting Method

Panelists used the Bookmark standard-setting method to establish cut scores for all subjects and grades. The Bookmark method is widely used in K–12 educational-assessment contexts. The Bookmark method uses panelists' review of collections of test items to generate cut scores (Cizek & Bunch, 2007). In this method, according to empirical item data (e.g., item response theory [IRT] item-parameter estimates), an ordered item booklet (OIB) displays items ranked

[7] Minor language change was implemented in 2022 on policy PLDs. The language was changed from "college and career readiness" to "postsecondary readiness," but the expectation for each achievement level remains the same.

from easiest to hardest. Panelists review the items in order and place a bookmark at the page in the OIB to indicate where they believe the *just-barely* examinee (i.e., minimally competent examinee or just-qualified candidate) has a specific probability (i.e., response probability) of answering the item correctly.

Taking advantage of IRT scaling, the Bookmark method places students and items on the same scale. According to the assumptions of the IRT model, a student's test score can provide a theoretically known probability for the student answering a dichotomous item (e.g., multiple-choice item) correctly. In the case of polytomously scored items (e.g., technology-enhanced items), responses are assigned a given score point. The student scores can be used to rank items.

According to Cizek and Bunch (2007), the Bookmark method is widely used for several reasons. First, from a practical standpoint, the method can be used for complex, mixed-format assessments, and panelists using the method can consider selected-response and technology-enhanced items together. Second, the method presents a relatively simple task for those participants who must make judgments regarding cut scores. Third, the Bookmark method is also comparatively easy for those participants who must implement the procedure. Finally, the method has certain psychometric advantages because of its basis in IRT analysis and its fidelity to test-construction techniques used during assessment development.

Given that KAP assessments are administered to a reasonably large population of students with an adequate number of assessment items across the range of performance, the Bookmark method was determined, in consultation with the Kansas State Department of Education (KSDE) and the Technical Advisory Committee, to be a reasonable method for establishing cut scores.

One key element of the Bookmark standard-setting method is the OIB. The OIB can contain both dichotomously scored items (e.g., multiple choice) and polytomously scored items (e.g., technology enhanced). Each dichotomously scored item appears once in the OIB in a location determined by the response probability (set as .67) and its IRT parameters. Each polytomously scored item appears several times in the OIB, once for each of its nonzero score points. Also, each page in the OIB corresponds to a score on the same scale.

VI.2.2. Procedures and Outcomes

The three standard settings (i.e., one for ELA and mathematics, one for science, and one for grade-10 mathematics) followed similar procedures but had slight changes to accommodate different timelines or location needs. The next sections summarize the standard-setting procedures and outcomes for these three standard-setting events.

VI.2.2.1. 2015 Standard Setting for English Language Arts and for Mathematics in Grades 3–8

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

Considering several aspects of panel diversity (e.g., ethnicity, gender, geographic area, teaching experience, and role), KSDE recruited 117 educators to be panelists for the standard-setting event. The *2015 KAP Technical Manual* describes the panelist recruitment and selection procedures, as well as their demographic characteristics. Panelists used policy PLDs (the *2015 KAP Technical Manual* describes the PLDs used for 2015 standard setting), just-barely student

statements (developed from grade-specific PLDs), and their experience teaching students to recommend cut scores.

Panelists were thoroughly trained before engaging in three standard-setting rounds. Before placing bookmarks, panelists

- took the operational test items
- defined the just-barely student at each performance level
- engaged in a practice activity
- described the knowledge and skills required to answer each test item
- completed a form to indicate their readiness for the standard– setting activities

Panelists completed three rounds of bookmark placements. For each round, panelists placed bookmarks for level 3, then for level 4, and finally for level 2. After Round 1, panelists reviewed their results and discussed table-level results. After Round 2, panelists reviewed table-level results, room-level results, and impact data. After the final round, panelists again reviewed room-level results and impact data. At the end of the standard-setting events, panelists completed the form to evaluate the standard-setting process and results. The evaluation results indicated participants felt (a) the opening training session was useful and helped prepare them for the standard-setting activities, (b) they had moderate to complete understanding of PLDs, (c) the results from previous rounds and materials provided during meeting were clear and useful, (d) the facilitators were helpful, and (e) impact results for cut scores at each level from Round 3 were reasonable.

After the standard-setting event and to ensure the reasonableness of cut scores across grades, AAI staff presented results from Round 3 to a policy-review panels. Approximately 40 educators participated in this phase. During the policy-review meeting, the facilitator introduced the Bookmark method procedures, provided the assertions and context for the standard-setting meeting, reviewed information about the PLDs, provided an overview of the steps in the standard-setting process, and discussed the materials that panelists used. The facilitator then presented impact data from the Round 3 bookmark-placement results. Policy-review panelists provided feedback about the process and the impact data. Then panelists considered the range of adjustment for the cut scores and recommended reasonable changes. At the end of the policy-review meeting, panelists provided feedback about the reasonableness, appropriateness, and defensibility of the cuts at level 2, level 3, and level 4. Policy-review panelists overwhelmingly reported that the cuts at levels 2, 3, and 4 were reasonable, appropriate, and defensible.

After KSDE presented the final cut scores from the policy review, the Kansas State Board of Education (the "State Board") approved the cut scores on September 8, 2015. Final cut scores approved by the State Board are in the *2015 KAP Technical Manual*.

VI.2.2.2. 2022 Grade-10 Mathematics Standard Setting

Standard setting for grade-10 mathematics occurred on July 29–30, 2022. Panelists collaborated during a virtual meeting to set cut scores for the new assessment. The next sections describe the panelists, PLDs, and the standard-setting procedure and outcomes for grade-10 mathematics. More-detailed information regarding material preparation, facilitator training, reliability and

validity evidence for the event, and materials used during standard setting are in the Mathematics Grade 10 Standard Setting Technical Report (Wang et al., 2022).

VI.2.2.2.1. Panelist Recruitment

For the KAP grade-10 mathematics standard-setting meeting, KSDE recruited panelists with experience teaching high school mathematics who were able and willing to participate in a completely virtual event. Overall, panelists formed a representative sample of Kansas public school educators. To obtain a large and diverse pool of applicants, KSDE began recruitment efforts in early 2022. KSDE sent a recruitment letter and interest survey via email distribution lists to curriculum leaders, test coordinators, and educators who provide mathematics instruction.

In total, KSDE recruited 39 educators as potential panelists for the event. KSDE asked all interested educators to complete an interest survey that requested basic demographic information and described the criteria for participation (see below). The survey asked educators to commit to up to 3.5 hours of advance training before the virtual standard-setting meetings and to attend 1.5 days of virtual standard-setting panel meetings on July 29–30, 2022.

KSDE identified several participation criteria prior to recruitment to ensure the selected panelists represented the following areas to the greatest extent possible:

- all 10 State Board of Education districts
- a cross-section of the state's large and small districts, rural and urban districts, socioeconomic status composition of districts
- a range of teaching experience (i.e., new and veteran teachers)
- experience teaching high school mathematics
- experience teaching students with disabilities
- experience teaching English learners
- experience teaching students from diverse backgrounds
- diversity in ethnicity/race and gender

To support the implementation of a virtual event, panelist-selection criteria also included:

- availability for a 1.5-day virtual event, plus approximately 3.5 hours of advance online training and activities via a Moodle course
- willingness to participate in the virtual event with honoraria or professional-development credit (if applicable)
- availability of a quiet and secure work area
- access to a desktop or laptop computer with internet connection (wired or wireless broadband: 4G, 5G, or LTE) and the following features:
    - participant's email
    - capability to participate in an online Zoom meeting, including:
        - speakers and a microphone
        - video capability

  - capability of running Kite® Student Portal software, including:
    - desktop or laptop running Windows 8.1 or 10, or macOS 10.14.5–11
    - one of the following browsers: Chrome, Edge, Firefox, or Safari
- access to any materials printed and mailed to panelists

KSDE selected 15 educators, who confirmed their intent and availability to complete the advance training and participate in the virtual panel meetings, and another 10 to serve as back-ups.

In the week before the standard-setting meeting, while the advance training took place, three educators declined participation because of personal or family emergency. Event staff contacted the back-up educators, they also indicated they had conflicts with the event dates and could not participate. Thus, 12 educators completed advance training and participated in the standard-setting process.

The 12 panelists who participated in the KAP grade-10 mathematics standard-setting event represented varying backgrounds, as summarized in Table VI-1. Among the panelists, most of them were female educators (n = 9, 75%), and all were White and non-Hispanic. Half of the selected panelists had 10 or more years of experience in high school mathematics (n = 6, 50%). Half (n = 6, 50%) of the panelists were from rural areas, and the other half were from urban or suburban areas. Approximately half the panelists had experience with students with disabilities (n = 7, 58%) and English learners (n = 6, 50%).

According to the 2017–2018 National Teacher and Principal Survey, 90.3% of Kansas public school teachers were White and non-Hispanic; 3% were Black and non-Hispanic; 2.5% were Hispanic, regardless of race (National Teacher and Principal Survey, 2020a); and more than three-fourths (i.e., 75.7%) were female (National Teacher and Principal Survey, 2020b).

The composition of the KAP grade-10 mathematics standard-setting panel (i.e., 75% female and 100% White) approximately represented the demographic characteristics of the Kansas public school teacher population.

Table VI-1. Panelist Demographic Characteristics for Grade-10 Mathematics Standard Setting (N = 12)

Characteristic	Group (*n*)	%
Gender		
Female	9	75
Male	2	17
Nonbinary	1	8
Area		
Rural	6	50
Suburban	5	42
Urban	1	8
High school mathematics experience		
1–5 years	5	42
6–10 years	1	8
11 or more years	6	50
Experience teaching students with disabilities		
Yes	7	58
No	5	42
Experience teaching English learners		
Yes	6	50
No	6	50
Role		
Classroom teacher	10	83
Classroom teacher and instructional coach	1	8
District staff	1	8

Among the 12 panelists participating in the standard-setting event, three of them performed the role of breakout-room leads during the Zoom meeting: they monitored the time and facilitated discussion during breakout sessions. Those three breakout-room leads expressed interest in the role before the event, and all were female with 4, 11, and 27 years of teaching experience in high school mathematics.

VI.2.2.2.2. Performance level Descriptors

As described in Section VI.1. State Adoption of Academic Achievement Standards for All Students, policy PLDs are the same across grades and subjects for the Kansas Standards. The KAP assessment uses policy PLDs to report student performance on score reports and to define the general expectations for student performance using four levels. The four levels categorize student performance and describe what students likely know and can do relative to the academic content standards.

- Level 1: A student at level 1 shows a limited ability to understand and use the skills and knowledge needed for postsecondary readiness.
- Level 2: A student at level 2 shows a basic ability to understand and use the skills and knowledge needed for postsecondary readiness.

- Level 3: A student at level 3 shows an effective ability to understand and use the skills and knowledge needed for postsecondary readiness.
- Level 4: A student at level 4 shows an excellent ability to understand and use the skills and knowledge needed for postsecondary readiness.

Students who achieve levels 3 and 4 are considered to have met the academic expectations of postsecondary readiness. For the purposes of standard setting, and to help set cut scores that differentiate student performance into four performance levels, ATLAS content-development staff created more-detailed descriptions of students' knowledge, skills, and abilities at each performance level for grade-10 mathematics. These more-detailed descriptions are referred to as *threshold PLDs* or *standard-setting PLDs*.

An important element for the Bookmark standard-setting procedure is a set of threshold PLDs. ATLAS content-development staff drafted the threshold PLDs for KAP grade-10 mathematics according to grade-specific PLDs and standards. The detailed descriptions of grade-specific PLDs can be found in Section III.2. Validity Evidence Based on Response Process. The draft threshold PLDs are intended to reflect the minimum key knowledge and skills of what students should know and be able to do to be categorized into each performance level. Threshold PLDs are intended to define the minimum policy-based and content-based expectations for each performance level. They are intended to assist standard-setting panelists in identifying the lowest-performing student who would qualify as meeting the expectations in a given performance level, that is, the student who would just barely meet the threshold (i.e., cut score) for being categorized in the given level. During standard setting, these students are referred to as *just-barely students*.

Panelists reviewed draft threshold PLDs before the standard-setting event. Then on the first day of the event, panelists discussed the draft threshold PLDs on the first day of the event, proposed edits or additions, and reached consensus on the final threshold PLDs, which were used throughout the entire standard-setting process.

VI.2.2.2.3. Standard-Setting Procedure

There were two main activities as part of the event: panelist advance training and assignments, and the virtual panel meeting. The purpose of the advance training was to let panelists become familiar with the grade-10 mathematics items and the Bookmark standard-setting method. For both activities, panelists used a Moodle course to develop, save, and deliver materials for the standard-setting activities.

VI.2.2.2.3.1. Panelist Advance Training and Assignments

All advance training took place one week before the meeting, from June 21–27, 2022. Advance training included a combination of synchronous and asynchronous activities conducted within the Moodle course (one exception is described in the next subsection) in advance of the virtual standard-setting meeting. The first activity provided self-paced training videos and a quiz for panelists, and then two assignments (i.e., taking the operational KAP grade-10 mathematics test and reviewing the draft threshold PLDs) concluded the advance training activities.

VI.2.2.2.3.1.1. Training Videos and Quiz

Event staff made the Moodle course available to the panelists on June 21, 2022. The course contained a series of training videos followed by a short quiz, a questionnaire about additional training needs, and a confidentiality form. The videos were available on demand and could be viewed any time during the training window. Panelists were required to watch all training videos before they could complete the online quiz. Panelists also signed a confidentiality agreement after completing the quiz. The following is a high-level outline of the training content.

- Video 1—Advance Training Orientation: A five-minute review of the panelist tasks in the advance training, including an overview of training videos and the two assignments.
- Video 2—KAP Grade-10 Mathematics Background, Test Design, and Policy PLDs: A 10-minute overview of the background for the grade-10 mathematics test and new cut scores required for reporting, grade-10 mathematics test design, item scoring, test scoring and reporting, and the policy PLDs for grade-10 mathematics.
- Video 3—Standard-Setting Overview: A 20-minute overview of the standard-setting meeting and the Bookmark method.
- Video 4—Step-by-Step Procedures of the Standard-Setting Meeting: A 20-minute overview of several activities that would take place during the virtual meeting before the bookmark-placement process began, an overview of the bookmark-placement process, and the individual steps of the process.
- Video 5—Meeting Attendees' Roles and Responsibilities: A 10-minute overview of panelists' roles during the standard-setting meeting, staff roles, materials to be used, the importance of materials security, and the consent to confidentiality.

Panelists completed a required six-question quiz covering critical points from the videos to ensure they completed all training videos. Panelists needed to answer all questions correctly before the standard-setting event. They were encouraged to review relevant parts of the training videos, if necessary, before retaking the quiz. Panelists were able to retake the quiz as many times as needed to score 100%. Ten out of 12 panelists achieved 100% in their first attempt. Two (16%) out of 12 panelists needed to take the quiz more than once; both of them retook the quiz and achieved 100%.

Panelists also responded to two open-ended items regarding any questions they still had from training or other areas in which they wanted additional information. No panelists had questions related to the training or the upcoming virtual standard-setting meeting.

In addition, the Moodle course included a Zoom guide, which described how to use Zoom tools and stay engaged in virtual meetings. Virtual office hours with ATLAS staff were available during designated times for panelists to log in to Zoom, test their software, practice using Zoom tools, and ask questions about Zoom. Moodle chat support was available for panelists to ask any questions about the training. ATLAS staff monitored the chat once each day during the training window.

VI.2.2.2.3.1.2. Assignments

Panelists completed two assignments within the Moodle course before the virtual standard-setting meeting. ATLAS staff conducted the first assignment synchronously through Zoom meetings, with two time slots available for panelists to choose from. The second assignment was

self-paced, to be completed after the first assignment but before the virtual standard-setting meeting.

- Assignment 1—Take the Operational Test: During a Zoom meeting, panelists accessed Kite Student Portal to take a proctored KAP grade-10 mathematics test. Two Zoom sessions were available: the afternoon of June 24, and the early evening of June 27. During the proctored test, ATLAS staff asked panelists to consider the items and test from students' perspective and to think about the kinds of knowledge and skills measured by each item. Panelists were instructed to submit their own questions through the Moodle forum; none were submitted.
- Assignment 2—Review Threshold Performance Level Descriptors: ATLAS staff instructed panelists to watch a short video that provided information about the purpose of threshold PLDs, how they were developed, and how to read the draft threshold PLD documents. Panelists reviewed the draft threshold PLDs, took notes related to the draft threshold PLDs, and prepared to discuss rationales for suggested changes on the first day of the virtual meeting. Instructions to panelists emphasized that draft threshold PLDs were derived from grade-specific PLDs that had been finalized and published, so significant changes to the PLDs were not expected.

VI.2.2.2.3.2. Virtual Standard-Setting Meeting

The 1.5-day meeting occurred through Zoom, and webcams were required. After the initial welcome session to review the standard-setting procedure and materials previously mailed to panelists, the panelists discussed threshold PLDs; completed a practice round, an OIB review, and three rounds of bookmarking; and ended the meeting with the articulation meeting and evaluation.

VI.2.2.2.3.2.1. Finalize Threshold Performance Level Descriptors

In a group, panelists discussed suggested edits and additions to draft threshold PLDs. The facilitator stressed the significance of this step and emphasized that conceptualizing what the just-barely students know and can do was critical to setting cut scores at each performance level. The facilitator reminded panelists that ATLAS content-development staff drafted threshold PLDs for KAP grade-10 mathematics according to grade-specific PLDs and standards. ATLAS staff and content experts at KSDE collaborated to develop the grade-specific PLDs, so major revisions were not expected. To establish panelists' clear understanding of the just-barely students, the panelists suggested and discussed edits to the threshold PLDs to differentiate performance expectations between performance levels. During the discussion, the facilitator reminded panelists to also consider the diversity of students in classrooms to ensure the inclusion of all student groups for finalizing the threshold PLD. ATLAS staff uploaded the finalized threshold PLDs to the Moodle course site for panelists' electronic access and for use during bookmarking rounds.

VI.2.2.2.3.2.2. Bookmark Practice

Panelists had an opportunity to practice and become familiar with the bookmark procedure. Using a practice OIB of seven items for grade-10 mathematics, a corresponding practice item map, and the final threshold PLDs, panelists completed the following steps:

- Panelists answered the question, "What does the student have to know and be able to do to answer each score point correctly?"
- Using the finalized level 3 (i.e., proficient level) threshold PLD, panelists answered the question, "Would 20 out of 30 just-barely level 3 students be able to answer each item correctly?"
- For polytomous items, panelists used the level 3 threshold PLD to answer the question, "Would 20 out of 30 just-barely level 3 students be able to earn this score point or higher?"
- Panelists used Zoom's Yes and No buttons to indicate their responses.
- Panelists discussed their rationales for their answers to these questions.

The practice round included only level 3 threshold PLD examples. When making the official placements, panelists would start with level 3 and then return to the first item in the OIB to continue the process for level 2 and then level 4.

VI.2.2.2.3.2.3. Review the Ordered Item Booklet

Panelists accessed the secure OIB in the Moodle course, and the facilitator led group discussion regarding panelists' perceptions of natural break points in the OIB for performance levels. During the discussion, panelists holistically reviewed items in the OIB to guard against one item excessively influencing the actual bookmarking process (i.e., panelists considered the entire group of items within which the cut score should be placed).

After reviewing the OIB, panelists participated in a readiness poll through Zoom to signal that no additional trainings were needed and they were ready to proceed with real bookmark placements.

VI.2.2.2.3.2.4. Setting Cut Scores

After panelists were comfortable with the rating procedure, they began the first round of item review and bookmark placements. The facilitator instructed panelists to refer to their materials organizer, which described the needed materials and where to find them. Panelists placed bookmarks during three rounds of ratings. Procedures in the next sections describe each of the three bookmark-placement rounds.

Round 1 Placement. To ensure the Round 1 bookmark placements were established independently, the facilitator instructed panelists to work alone and avoid discussion with others. Panelists divided into three Zoom breakout rooms and worked individually using the OIB, item-map table, just-barely student definition, and colored bookmark-placement form.

Starting with the first item in the OIB, panelists reviewed each item individually. They placed bookmarks where two-thirds of the just-barely level 3 students would be able to answer the item correctly (or obtain the score point for polytomous items). Panelists reviewed a few items after the bookmarked item to make sure that just-barely level 3 students would not be able to answer subsequent items correctly. After panelists placed their bookmarks for the just-barely level 3 students, they returned to the first item and followed the same procedure for the just-barely level 2 students. They then placed their bookmarks for just-barely level 4 students.

After panelists finished their bookmark placements for the three cuts, they wrote their bookmarks on the cardstock form provided by mail. Then, they submitted their placements in the Google Form using a preassigned panelist ID number.

Round 1 Results and Discussion. All panelists came back to the main Zoom meeting. The facilitator displayed (i.e., shared their screen in Zoom) the Round 1 summary results derived from panelists' bookmark placements. The results included the three different breakout groups and the whole panel. Table VI-2 and Figure VI-1 provide example results. The facilitator pointed out that the bar chart (Figure VI-1) was intended to show all individual placements for the panel. Panelists compared their own results with those of the other panelists in the same breakout room and in the whole panel, and then were asked to consider three questions:

- Am I relatively strict or lenient in relation to others in the same breakout room group?
- Am I relatively strict or lenient in relation to others in the whole panel?
- Am I consistently strict or lenient across all three levels?

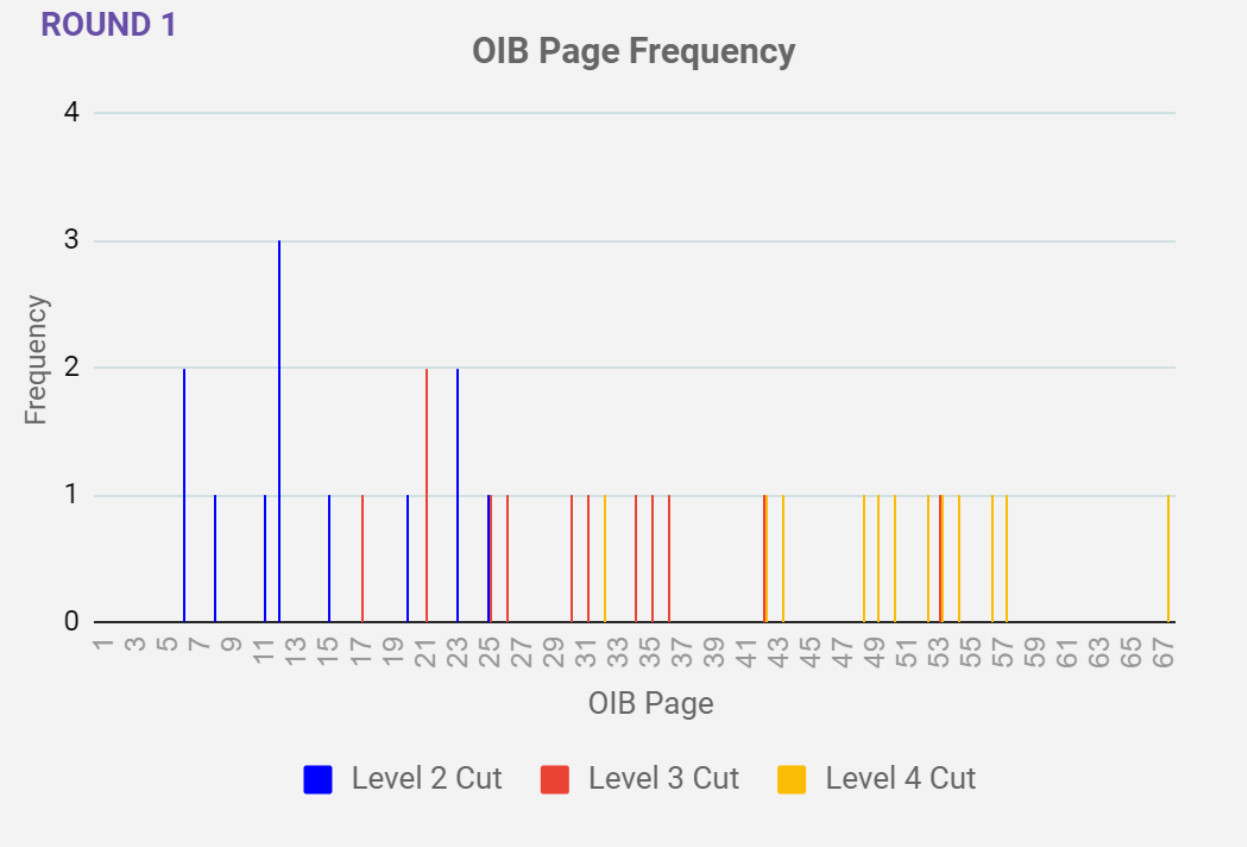
The facilitator placed the summary results in the Moodle course so panelists could download them to their computers and use them during discussion. Panelists went into breakout rooms to review and discuss Round 1 summary results. Throughout the discussion, the facilitator encouraged panelists to consider other perspectives, describe their thoughts, and provide content rationale to their breakout-room colleagues. Some panelists offered rationales for higher or lower bookmark placements. The group discussed panelists' thoughts about items that fell between minimum and maximum bookmarks for each level. While discussing rationales, panelists shared the perspectives and experiences that contributed to their expectations for student performance.

Table VI-2. Grade-10 Mathematics Example Summary of Results for Bookmark Placements

	Performance Level Cuts		
	Level 2	Level 3	Level 4
OIB minimum	6	17	32
OIB median	12	30	51
OIB maximum	25	53	67

Note. OIB = ordered item booklet.

Figure VI-1. Example Frequency of Round 1 OIB Page Numbers With Bookmarks for Grade-10 Mathematics

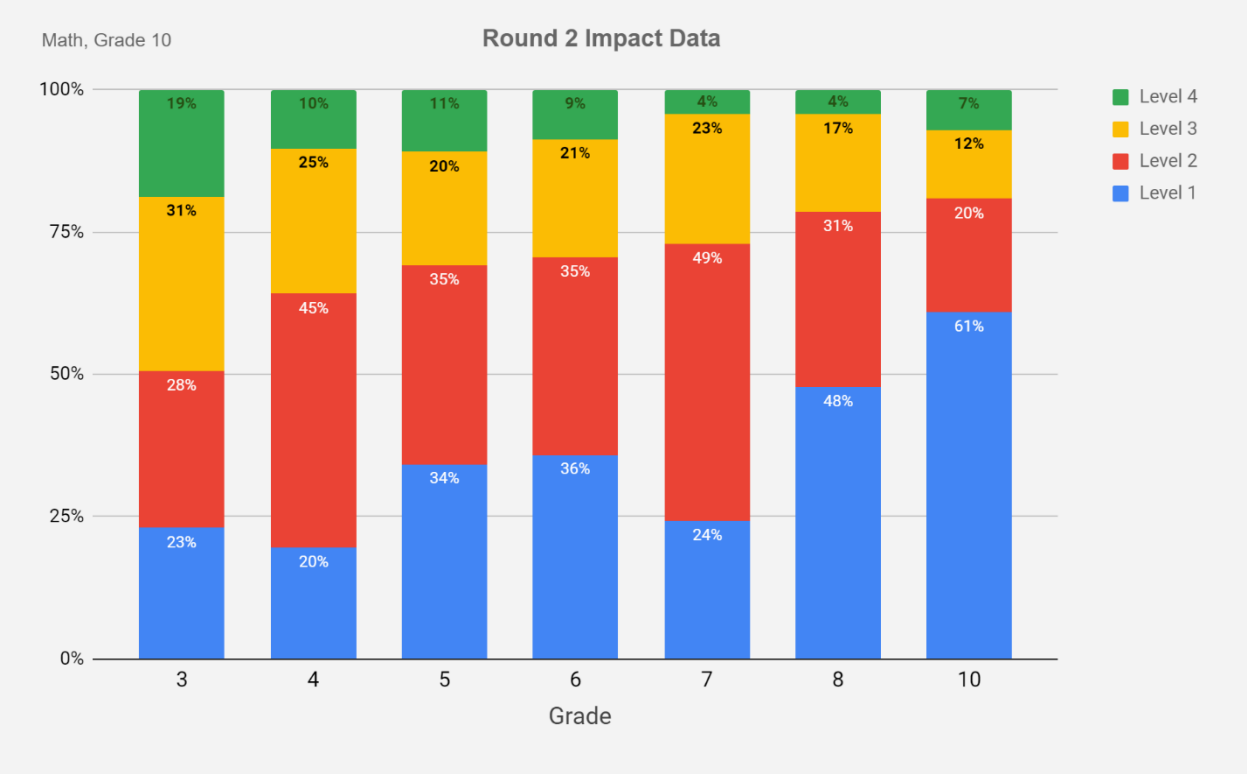


Round 2 Placement. After discussion, the breakout-room groups returned to the main Zoom meeting as one panel. Panelists repeated the Round 1 procedures, considering Round 1 feedback results and panel discussions. Additionally, the facilitator informed panelists that they did not need to conform to the Round 1 median when they placed the bookmarks for Round 2, as new information might influence the direction of their placements.

At the end of the second round, panelists submitted their placements through a Google Form. For the second round, ATLAS staff prepared a whole-panel summary table and chart in the same format used in the first round. The data tool also generated impact data after this round.

Round 2 Results and Discussion. The facilitator shared their computer screen in Zoom to show the results of the second round. The facilitator also showed the panelists impact data as a stacked bar chart, which displayed the percentage of students in each performance level according to the cut scores derived from the panel's median bookmark placement from the second round (see Figure VI-2); it also included real student-performance data from the 2022 KAP administration. The results also included impact data from grades 3–8 to provide a context for panelists as they compared grade-10 achievement expectations with the established achievement expectations in grades 3–8 (i.e., the consistency of performance expectations across grades).

Figure VI-2. Round 2 Impact Data for Grade-10 Mathematics



Panelists again compared their own bookmarks with those of other panelists and considered whether they were consistently lenient or strict in relation to others. The facilitator guided panelists to think about three questions and to share their thoughts and rationales with the group:

- What is the range of the bookmark placements?
- How has the range for Round 2 changed in comparison to Round 1?
- How does my bookmark placement compare to the panel's average placement?

Round 3 Placement. The facilitator summarized the tasks of the third round, emphasizing the third round as the final opportunity for panelists to revise their bookmark placements. Median placements from this round would be the final, panel-recommended cut score and would be used during articulation.

The facilitator advised panelists to use all available information to guide their decisions: individual and median bookmark placements over the two rounds, threshold PLDs, notes from reviewing the OIB, impact data, and the input of their colleagues through discussion. Panelists entered their placements into a Google Form after recording their final bookmark on the cardstock form. The same summary table and chart prepared for the second round were also prepared for the third round and presented during articulation.

VI.2.2.2.3.2.5. Vertical Articulation Procedure

After the standard-setting panel meetings ended, all panelists served on the articulation panel. The goal of articulation is to align grade-10 mathematics cut scores with grades 3–8 cut scores to avoid unintended or inappropriate reversals of performance expectations for grade 10 compared

to other grades, as the cut scores for other grades were set in 2015. In this articulation, only the grade-10 cut scores could be adjusted.

Panelists first completed training on articulation. The training covered the purpose of articulation, the responsibilities of panelists during articulation, and what to consider during the articulation process. After the training, the facilitator presented the impact data for all grades for mathematics, which were derived from the panel's median bookmark placement from Round 3 for grade 10. After reviewing the impact data, panelists shared the expected student's achievement-level distribution across grades based on their experiences. Because panelists were grounded in their content rationales for their ratings in Round 3 and the cross-grade impact data shown after Round 2, no adjustments to the cut scores were made during articulation.

VI.2.2.2.3.2.6. Evaluation

After articulation, panelists completed the cut-score evaluation form and the standard-setting process-evaluation form. Then, the facilitator and the assistant director of assessment from KSDE thanked the panelists for participating in the virtual standard-setting meeting, after which the meeting adjourned. Discussions regarding the cut-score and standard-setting evaluation results are in Section VI.2.2.2.5. Panelist Evaluation.

VI.2.2.2.4. Standard-Setting Results

Table VI-3 shows the median OIB page number by rounds. The scale scores associated with the Round 3 median OIB page number among panelists' rating are the final panel-recommended cut scores. These scale-score cuts are 462 for Level 2, 568 for Level 3, and 695 for Level 4. Note that these scale-score cuts are expressed in a temporary standard-setting scale and do not reflect the final reporting scale for grade-10 mathematics

Table VI-3. Median Ordered Item Booklet Page by Round for Grade-10 Mathematics

Round	Level 2 cut	Level 3 cut	Level 4 cut
1	12	30	51
2	12	27	48
3	8	23	46

ATLAS staff calculated the percentages of students in each performance level (i.e., impact data) from the scale-score cuts recommended by the panel shown in Table VI-3. The impact data can be found in Section IV.4.3.2. Operational Test Results.

VI.2.2.2.5. Panelist Evaluation

As described in Section VI.2.2.2.3.2.6. Evaluation, panelists completed both the cut-score evaluation and the evaluation of the standard-setting process. The cut-score evaluation implemented three rounds of bookmark placements and articulation. The next subsections summarize the questions and results from evaluation.

VI.2.2.2.5.1. Panelist Cut-Score Evaluations

The cut-score evaluation consisted of two main sections: panelists' perceptions of influential factors in their bookmark placements as well as articulation, and their perceptions of the panel

bookmark-placement results. For each performance level, the evaluation asked panelists to first indicate the importance of each influential factor as they placed their bookmarks: *not important*, *slightly important*, *moderately important*, *very important*, or *not applicable*. The evaluation then asked panelists to indicate the extent of their agreement or disagreement with statements related to the bookmark-placement results: *strongly disagree*, *disagree*, *somewhat disagree*, *somewhat agree*, *agree*, or *strongly agree*.

Panelist-evaluation results regarding influential factors varied for each round. Panelists chose their experience with students at this grade as the most influential factor for Round 1; 83% of panelists rated this very important. Panelists chose their experience with the 2017 Kansas Standards as the second-most influential factor, with 75% of panelists rating it very important. These two highly rated elements in terms of importance were also deemed very important by the majority of the panelists (75%) in Round 2. Panelists rated group discussion as an equally influential factor in Round 2, as 75% of panelists rated it very important. This pattern persisted in Round 3; 75% of panelists considered their experience with students at this grade and their experience with 2017 Kansas Standards to be very important, and 67% of panelists rated group discussions as very important. For articulation, 83% of the panelists said group discussion was very important, and 75% of the panelists rated policy PLDs and the threshold PLDs and the experience with students at this grade as very important. The least influential factor across all three rounds, other than articulation, was the bookmark placements of other panelists during prior rounds; 33% of panelists rated it slightly important in Round 2, as did 42% in Round 3.

Responses regarding bookmark placement were generally positive. The percentage of panelists who agreed or strongly agreed to all four statements measuring clarity and usefulness of summary panel results from Rounds 1 and 2 (ranging from approximately 75%–100%) indicated that the tables and graphs were clear and useful. Additionally, all panelists agreed or strongly agreed that the impact results for level 2, level 3, and level 4 were reasonable; that the cut score for each level was appropriate according to the policy PLDs and threshold PLDs; and that the cut score for each level was defensible because of panelists' adherence to procedures.

VI.2.2.2.5.2. Standard-Setting-Process Evaluation

The standard-setting-process evaluation included questions about panelists' perceptions of various aspects of the advance training activities and virtual panel meetings. The questions were organized into nine strands and used various Likert rating scales.

First, the majority of panelists (83%–100%) agreed or strongly agreed with statements regarding the effectiveness, clarity, and organization of the advance training and assignments. For example, most panelists (83%) agreed or strongly agreed that the advance training and activities helped them prepare for the standard-setting event and clearly explained the meeting procedures. Most panelists (83%–100%) agreed or strongly agreed with statements regarding the effectiveness, importance, and organization of the welcome and orientation sessions (100%), the group discussions (83%), and the practice sessions (100%). Most panelists said that the amount of time used for advance training and assignments was about right (67%). Similarly, most said that the amount of time used for orientation and the additional training during the standard-setting meeting was about right (75%).

Also, nearly all panelists (85%–100%) agreed or strongly agreed with statements regarding the bookmark-placement activities. For example, all panelists agreed or strongly agreed that the

bookmark-placement forms (both cardstock paper and Google Form) were easy to understand and use, that the expectations for each round of bookmark placement were clear, and that they made their bookmark placements on their own during the independent bookmark-placement process. The only statement that 25% of panelists either disagreed or somewhat disagreed with was regarding the amount of time needed to complete each round, as the panelists felt the time was more than sufficient to complete the activities. Most panelists (83%–100%) agreed or strongly agreed with statements regarding group discussion. For example, all panelists agreed or strongly agreed that everyone had equal opportunity to contribute ideas and opinions, and that discussions after each round of ratings were helpful. The lowest ratings (83% agreed or strongly agreed) were in response to the statement, "The quality of the group discussion was not negatively impacted by the virtual setting of the meeting," which may indicate in-person discussion is still preferred by some panelists.

Furthermore, all panelists agreed or strongly agreed that the training of articulation was helpful, they had equal chances to contribute thoughts, and the time provided was adequate. Additionally, nearly all panelists (92%) agreed or strongly agreed that the expectation for articulation was clear and that they understood the expectation.

Regarding materials, technology, and staff, panelists also provided positive feedback. While only 17% of panelists found the Zoom 101 reference guide moderately useful or very useful, most panelists (66%–92%) found the other materials moderately useful or very useful during the standard-setting process. Although 17% of panelists experienced problems logging in to Student Portal, the vast majority of panelists (92%–100%) agreed or strongly agreed that the other technology features (including features in Moodle, Zoom, and Google Forms) were effective or easy to use. Nearly all panelists (92%–100%) believed the panel's lead facilitator, the Zoom host, and the content specialist were moderately helpful or very helpful; 75% panelists rated other staff very helpful, and 25% said this role was not applicable during the event.

VI.2.2.3. 2017 Science Standard Setting

AAI conducted a standard-setting event for science using the Bookmark method during a workshop held at a school in Topeka on June 20–21, 2017. The standard-setting event included a training session and three rounds of the Bookmark procedure for each grade and subject-area test. The main goal of the science standard setting was to establish three cut scores to differentiate four performance levels for the assessment. The next sections describe the panelists who participated, PLDs, and the standard-setting procedure and outcomes for science. More-detailed information regarding preparation milestones, reliability and validity evidence for the event, and materials used during standard setting are in the Science Standard Setting Technical Report (AAI, 2017).

VI.2.2.3.1. Panelist Recruitment

The selection and training of the standard-setting panelists was crucial to the success of the standard-setting event. Considering several aspects of panel diversity (e.g., ethnicity, gender, geographic area, teaching experience, 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 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 had not participated 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

- all 10 State Board of Education districts
- priority/focus schools
- a cross-section of state large or small districts, rural or urban districts, and socioeconomic composition of districts
- a range of length of teaching experience (i.e., new or veteran teacher)

Panelists were recruited with the goal of having at least 12 panelists participating for each grade. Because of panelist attrition shortly before the event, grade 8 included 11, rather than 12, panelists. The grade 5 had 13 panelists, and grade 11 had 15 panelists. Table VI-4 summarizes the demographic characteristics of all panelists by grade. Nearly half of the panelists were from rural areas, and the other half were from urban or suburban areas. The average number of years of experience was approximately 13.5. Most of the panelists were White female educators. According to the demographic survey summarized in Section VI.2.2.2.1 Panelist Recruitment in the grade-10 mathematics standard setting, the composition of the KAP science standard-setting panel (i.e., 75% female and 88% White) approximately represented the demographic characteristics of the public school teacher population in Kansas.

Table VI-4. Demographic Characteristics of Panelists for Science Standard Setting, by Grade

Subgroups	Grade 5 % (*N* = 13)	Grade 8 % (*N* = 11)	Grade 11 % (*N* = 15)
Gender			
Female	85	91	63
Male	15	9	38
Race			
White	85	91	94
Others	15	9	6
Area			
Rural	38	31	63
Suburban	31	38	19
Urban	31	15	19
Teaching experience (years)			
M	9.5	14.5	16.2
SD	5.4	7.7	6.6

VI.2.2.3.2. Performance Level Descriptors

PLDs are the guiding performance standards when setting cut scores. The creation of grade-specific PLDs for science began with KSDE and AAI 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 C). KSDE staff and Kansas educators reviewed and approved the grade-specific PLDs for all four levels before the standard-setting workshop.

VI.2.2.3.3. Standard-Setting Procedure

During the standard-setting meeting, panelists were separated into three panels, one for each grade. Each grade sat at three tables, and each table had three to five panelists. For each grade, the standard-setting procedures were steered by a lead facilitator and three table leads recruited by AAI from among the selected panelists. Both KSDE assessment personnel and AAI content-team members were available during the workshop to address policy-related or content-related questions. A description of the workshop structure follows.

- June 20
  - Completed the training session
  - Completed 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 the 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.2.2.3.3.1. Training Session

At the start of the meeting, panelists completed a participant survey form and signed a confidentiality form. The survey collected panelists' biographical data to contribute to the documentation of the procedural validity of the standard-setting process (Hambleton et al., 2012; Pitoniak & Morgan, 2012). Then, AAI staff conducted a large-group training to address general topics, which included an overview of the science assessment and an introduction to the concept of cut scores. AAI 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 panelists' questions. Room facilitators also answered standard-setting-related questions generated by panelists at their tables; policy-related questions were directed to KSDE staff.

VI.2.2.3.3.2. Taking the Science Assessment

To provide a frame of reference for considering student performances, the panelists took the science assessment in a shorter timeframe than is used operationally. The panelists took the assessment in the Kite system using Chromebooks supplied by the school. Students use Chromebooks during operational testing, so their use by 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 students did for their operational test. The panelists were given 45 minutes to finish the test and were encouraged to think about how students experienced the items. After becoming familiar with item and test difficulty while taking the test, the panelists discussed item and test difficulty.

VI.2.2.3.3.3. 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, that is, students whose scores would put them just barely in the level. The student score at the bottom of the level defines that cut score. Panelists were guided to use the just-barely worksheets to help them define the performances of students in this area. They used the just-barely worksheets to answer the question, "What knowledge and skills does this student have that a student who is at the top of the lower adjacent level does not have?" They started working individually and then had a group discussion. A

description of just-barely performances for each achievement level was approved by panelists at the end of this activity.

VI.2.2.3.3.4. Bookmark Practice

The purpose of this practice is for panelists to become familiar with the bookmark procedures. Using a practice OIB of 10 items, the practice item-map table, and the practice item-dot-plot sheet, the panelists reviewed the practice items and considered two 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 items preceding it?

Panelists discussed the knowledge and skills required to correctly answer the first item on the practice OIB, guided by these two questions. They referred to their just-barely student list for level 3 and asked themselves if two-thirds of the just-barely level 3 students would be able to answer this item correctly. For items with 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 most panelists thought those students would earn that point, they proceeded to the second item. Panelists continued this process until they agreed that a just-barely level 3 student would not be able to earn the score point; they then placed the bookmark on that item. They repeated this process to place the level 4 and level 2 bookmarks.

VI.2.2.3.3.5. 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 the same two 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 items preceding it?

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

VI.2.2.3.3.6. Setting Cut Score: Round 1

Panelists obtained item-map tables, item dot plots, and OIBs. Before the Round 1 rating, all panelists completed the readiness form indicating they were ready to begin actual bookmarking. They started with the first item and continued until they felt they had reached the point at which two-thirds of the just-barely level 3 students would not be able to answer the item correctly. The panelists placed a bookmark on that page and recorded the page number on the placement form. After they placed their bookmarks, the panelists were reminded to consider a few items after the marked item to be sure they had their bookmarks in the right place. They also were reminded to consider the Kansas Standards for Science, PLDs, just-barely lists, and their notes about knowledge and skills. After placing the level 3 bookmark, panelists continued to place level 4 bookmarks and then level 2 bookmarks at the appropriate places.

There were three rules for panelists' bookmark placements.

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 in the booklet by difficulty, from easiest to hardest. The level 2 bookmark page should appear earliest in the booklet, and the level 4 bookmark page should appear latest in the booklet among the three levels of bookmark pages.

Panelists completed this work independently. After completing their bookmark placements, they submitted their bookmark placements to the facilitator and completed the evaluation form on training, practice, just-barely student activity, and influential factors for Round 1.

VI.2.2.3.3.7. Setting Cut Score: Round 2

Each table as well as the whole panel discussed Round 1. Panelists for each grade first reviewed and discussed Round 1 bookmark summary results. They were asked to consider several questions.

- How tough or easy were 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?

Panelists then considered how their individual ratings compared to others' ratings, using several guiding questions.

- 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 that point?

Panelists were encouraged to use information from Round 1 results to inform themselves and to give themselves the chance to reconsider ratings. They were instructed to consider the Kansas Standards for Science, PLDs, just-barely attributes of students, and their item knowledge and skill notes when placing bookmarks. The same procedures for placing bookmarks used in Round 1 were used again. Panelists were told that they could change their bookmark placements or keep them the same. After completing bookmark placement, panelists submitted their bookmark-placement forms to their facilitators and completed the evaluation form on influential factors for Round 2.

VI.2.2.3.3.8. 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 that point, considering the following questions:

- 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 aspect for panelists to consider was the impact data (i.e., performance-level distribution) from Round 2 results. With information provided by the impact data, panelists learned 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 impct data based on their Round 2 recommendations. Finally, a graph showing how Kansas students fared on the National Assessment of Educational Progress (NAEP) science assessments in 2015 on the nearest grade 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 results; panelists for grades 8 and 11 saw grade-8 NAEP results.

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 can 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 assimilated all the information and placed their best and final bookmarks. Panelists submitted their bookmark placements and completed the remaining sections of the evaluation form.

VI.2.2.3.4. Standard-Setting Results

Table VI-5 presents the median bookmark placements among panelists for grades 5, 8, and 11 for each round. The scale scores associated with the Round 3 median OIB page number among panelists' rating are the panel-recommended cut scores from the standard-setting meeting.

Table VI-5. Rounds 1–3 Median Bookmark Placements by Grade for Science

Round	Grade 5			Grade 8			Grade 11		
	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
3	9	22	37.0	14	28.0	47.0	10	25	40

VI.2.2.3.5. Articulation

The objective of the articulation meeting is to help ensure the reasonableness of cut scores across grades. Table leaders from grade panels were recruited for this meeting. There were three articulation panelists from grade 11, three from grade 8, and two from grade 5.

The articulation meeting started with articulation training to help 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
- expectation on the level cut consistency across grades
- articulation procedure
- standard error of judgment
- reasonable level cut-score adjusting range

During the articulation, the articulation leader first presented the Round 3 level cuts, the impact data of all grades, and the reasonable ranges within which the level cuts could be adjusted (cut scores ± 1 × conditional standard error of measurement). Panelists then discussed these results using the following questions:

- Are there differences between the impact data and what you expected the impact would look like based on your recommended cut scores?
- If there are differences, why do you think the impact data does not match your expectations?

The articulation leader answered panelists' questions about the articulation before they started discussing the articulation as a group. 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 showed panelists the change of the impact data after adjusting the level cut score. Finally, panelists completed the evaluation form of the articulation section.

The final cut scores recommended by the articulation panels are described in Section IV.4.2.2. Scale-Transformation Constant. The State Board approved these cut scores in fall 2017.

VI.2.2.2.6. Panelist Evaluation

Throughout standard setting and articulation, panelists evaluated the standard-setting process, the articulation process, and the reasonableness of cut scores. Regarding the standard-setting process, most panelists agreed that they had a clear expectation of the process, made their ratings independently, and had enough time to finish rating. Regarding the articulation process, panelists generally perceived the processes and information in the articulation meeting to be clear. Table VI-6 summarizes panelists' responses to questions on the evaluation form about the results of cut scores. On average, they agreed that the cut scores were reasonable according to the impact data and PLDs. Grade 11 showed less agreement compare to other grades, which may have been caused by the more-diverse panelist composite.

Table VI-6. Summary From Evaluation Survey of Panelists' Perceptions of Cut-Score Results for Science Standard Setting

Statement	Grade mean [a]		
	5	8	11
Grade group results for 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 group results for 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 group results for 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

Note. [a] Likert scale score: 1 = strongly disagree, 6 = strongly agree.

VI.3. Challenging and Aligned Academic Achievement Standards

Educators set the KAP's academic achievement standards to align with the state content standards (i.e., Kansas Standards). First, ATLAS content experts worked alongside KSDE staff to define grade-specific PLDs. The grade-specific PLDs were at the standard level for ELA, at the cluster level for mathematics, and at the target level for science. The iterative process ended when both sides agreed that the expected performance adhered to content standards, as well as to cognitive demands, and that the overall expectation properly reflected the rigor of the Kansas Standards. Then, the grade-specific PLDs were presented to Kansas educators for review and approval. As described in Section VI.2. Achievement Standard Setting, grade-specific PLDs are the foundation for developing threshold PLDs, which is the key documentation to help educators determine the academic achievement standards.

VI.4. Reporting

For each tested grade and subject, the KAP assessment provides separate score reports to students, schools, and districts. Examples of a KAP student score report, KAP school report, and district report are in Appendix D. These reports include students' overall and subscore performances. Score reports present the results 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 results. As stated in Petersen et al. (1989), providing score interpretations in score reports can minimize misinterpretations and unwarranted inferences. Helping readers understand the meaning of the statistics is as important as reporting the values.

Although these reports are intended for different groups (e.g., students, schools, 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. The uniform design eases educators' burden of review and helps them explain score reports to parents.

Moreover, the state added language to all score reports for the 2022 administration to remind students, parents, and educators that learning conditions and student performance may still have been affected by COVID-19. This caveat states, "When interpreting KAP results, please take into consideration other measures of student achievement. Also, consider how the conditions for learning, which may have been disrupted by the pandemic, may influence performance" (Appendix D).

VI.4.1. Student Reports

Samples of a grade-10 mathematics student report and a science student report are in Appendix D. In the student report, a student's performance level is given immediately after student identifiers so that the performance level is the first information presented. Next are the student's scale 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 grade-specific PLDs that describe what the student should be able to do. Score meters report the medians of school, district, and state performances. The report shows the median 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 next section shows the standard error of measurement (*SE*) of scale scores and *SE*s of school, district, and state median scores. The *SE* reported on student scores is the conditional standard error of measurement (CSEM) derived from the IRT scaling model. The CSEM indicates how much a student score might vary if the student took many equivalent versions of the test. The *SE* of group scores (i.e., school, district, state) is the *SE* of the median account for sampling error but not for measurement error. The *SE* of the median is computed using Equation VI-1.

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

where $SE_{median(x)}$ is the *SE* of the median of the group scores, $S_{\bar{x}}$ is the standard deviation of the group's observed scores, and N is the number of students in the group. The final section of the student report shows the overall policy PLD for each performance level.

The second page reports the student's performance by subscores. This information indicates strengths and weaknesses in different domains or clusters for ELA and mathematics or in claims for science. Each subscore represents a group of test items that assesses related skills. For the reported subscores for different grades and subject, refer to Section IV.1.4. Subscore Reliability.

VI.4.2. 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 performance level (level 1 through level 4). Samples of mathematics school and district reports are in Appendix D.

School reports provide summary information of a subject by grade. On the first page, bar graphs indicate a school's median scale scores for each grade, along with scores of the school's district and state overall performances. The report includes district and state median scale scores as a reference so schools can interpret their standing. The standard errors 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, the school report provides district and state results 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 aggregated performance for different subscores and descriptions of each subscore category. The percentage of students in each subscore determines the aggregated rating of each subscore. If a rating is obtained by more than 50% of students in the school, then the rating is the aggregated rating. Section IV.1.4. Subscore Reliability describes the calculation of students' subscore rating.

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

When group counts are very small, individual students may be identified through demographic information, even on group summary reports. For a school or district with fewer than 10 students, the school or district report is not available for KAP.

VI.4.3. Reporting Timeline

The KAP testing window ended on April 28, 2022. One week later, KSDE began a review of KAP ELA, mathematics (except grade 10), and science score reports. After KSDE approved the score reports, districts and parents were given access to them.

For grade-10 mathematics, AAI and KSDE staff presented cut scores to the State Board after the 2022 standard-setting meeting. The State Board approved the cut scores on August 9, 2022. KSDE reviewed the reports for grade-10 mathematics on August 12, 2022. After KSDE approved the score reports, districts and parents were given access to them.

VI.4.4. Interpretive Guides

To help educators and parents interpret KAP results, the *KAP Educator Guide* and the *KAP Parent Guide* with a Spanish version are available on the KAP website so that educators and parents can access them easily. Both guides include a letter from Dr. Randy Watson, Kansas Commissioner of Education; an overview of test purposes, content, and format; descriptions of the KAP scoring process; suggestions for how to use test scores and improve KAP scores; and an explanation of the different information presented on the score reports.

References

Achievement and Assessment Institute. (2017). *KAP science standard setting technical report*. University of Kansas, Achievement and Assessment Institute (AAI).

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin*, *102*, 3–27. https://psycnet.apa.org/doi/10.1037/0033-2909.102.1.3

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

Anderson, J. R. (1992). Automaticity and the ACT theory. *The American Journal of Psychology*, *105*(2), 165–180. https://doi.org/10.2307/1423026

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355. https://psycnet.apa.org/doi/10.1037/0003-066X.51.4.355

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: contemporary methods. *Educational Measurement, Issues and Practice*, *23*(4), 31–50. https://doi.org/10.1111/j.1745-3992.2004.tb00166.x

Egan, K., & Davidson, A. (2022a). *Alignment study for Kansas Assessment Program, mathematics grade 10*. EdMetric LLC.

Egan, K., & Davidson, A. (2022b). *Alignment study for Kansas Assessment Program, grade 5, middle school, and high school science*. EdMetric LLC.

Forte, E. (2013, June 19–22). *Next generation alignment approaches: Needs and promising directions.* Paper presented at the National Conference on Student Assessment, Baltimore, Maryland, United States.

Forte, E. (2016). *Evaluating alignment in large-scale standards-based assessment systems* [White paper]. Technical Issues in Large Scale Assessment State Collaborative on Assessments and Student Standards of the Council of Chief State School Officers. https://edcount.com/wp-content/uploads/2019/05/ccsso_tilsa_forte_evaluating_alignment_2017.pdf

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. https://www.jstor.org/stable/1434586

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 37–50. https://doi.org/10.1207/s15324818ame0201_3

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). Routledge.

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation*, *17*(6), 497–509. https://doi.org/10.1080/13803611.2011.632668

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. https://doi.org/10.1207/S15324818AME1404_2

Kan. Stat. Ann. §72-5170 (2020). https://www.ksrevisor.org/statutes/chapters/ch72/072_051_0070.html

Kansas State Department of Education. (2017). *2017 Kansas mathematics standards: Grades K-12*.

Li, F., Cohen, A., & Shen, L. (2012). Investigating the effect of item position in computer-based tests. *Journal of Educational Measurement*, *49*(4), 362–379. http://www.jstor.org/stable/23353876

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. https://www.jstor.org/stable/1435147

Logan, G. D. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology*, *39*(2), 367–386. https://psycnet.apa.org/doi/10.1037/h0080066

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scores items. *Journal of Educational Measurement*, 30, 107–122. https://doi.org/10.1111/j.1745-3984.1993.tb01069.x

National Teacher and Principal Survey. (2020a). *Percentage distribution of public school teachers, by race/ethnicity and state: 2017–18* [Data set]. U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences. https://nces.ed.gov/surveys/ntps/tables/ntps1718_fltable01_t1s.asp

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). 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). Routledge.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1988). *Numerical recipes*. Cambridge University Press.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Lawrence Erlbaum Associates.

Wang, W., Swift, D., & Yu, H. (2022). *2022 KAP mathematics grade 10 standard setting technical report*. University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS).

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6) (ED 414305). ERIC. https://files.eric.ed.gov/fulltext/ED414305.pdf

Webb, N. L. (1999). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Council of Chief State School Officers.

Wine M., & Hoffman A. (2020). *Rigorous test development (RTD): Theory of the item* [White paper]. AleDev Research/ResearchGate.

https://www.researchgate.net/publication/362161122_Rigorous_Test_Development_RTD_Theory_of_the_Item

Wine, M. & Hoffman, A. (2022). *RTD approach to using Norman Webb's depth of knowledge (DOK) typology of cognitive complexity*. ResearchGate. 10.13140/RG.2.2.13393.61280

Appendix A: Blueprints by Grade

Table A-1. English Language Arts Blueprint Across All Grades

Grade	Domain or subdomain	Clusters	Depth of knowledge	% of items
3–10	Writing	Text types and purposes Language in writing	2	35–40
	Reading: literacy	Key ideas and details Craft and structure Language in reading Integration of knowledge and ideas	2, 3	30–35
	Reading: information	Key ideas and details Craft and structure Language in reading Integration of knowledge and ideas	2, 3	30–35

Table A-2. Mathematics Blueprint by Grade

Grade	Classification	Domain/ Conceptual categories	Depth of knowledge	% of items
3	Skills and concepts	Operations and algebraic thinking Numbers and operations with fractions Measurement and data Geometry	1, 2	75–88
4	Skills and concepts	Operations and algebraic thinking Number and operations in base ten Numbers and operations with fractions Measurement and data	1, 2	75–88
5	Skills and concepts	Number and operations in base ten Numbers and operations with fractions Measurement and data	1, 2	75–88
6	Skills and concepts	Ratios and proportional relationships The number system Expressions and equations Geometry Statistics and probability	1, 2	75–88
7	Skills and concepts	Ratios and proportional relationships The number system Expressions and equations Geometry Statistics and probability	1, 2	75–88
8	Skills and concepts	Expressions and equations Functions Geometry	1, 2	75–88
10	Skills and concepts	Number and quantity and algebra Functions Geometry Statistics and probability	1, 2	75–88

Grade	Classification	Domain/ Conceptual categories	Depth of knowledge	% of items
3–10	Strategic thinking and reasoning	Problem-solving and modeling Communication reasoning	2, 3	12-25

Table A-3. Science Blueprint by Grade

Grade	Claim	Target	Depth of knowledge	% of items
5	Physical science	Structure and properties of matter Engineering design in physical science	2, 3	27–33
	Life science	Matter and energy in organisms and ecosystems Engineering design in life science	2, 3	34–40
	Earth and space science	Earth's systems Space systems Engineering design in Earth and space science	2, 3	27–33
8	Physical science	Structure and properties of matter Chemical reactions Forces and interactions Energy Waves and electromagnetic radiation Engineering design in physical science	2, 3	27–33
	Life science	Structure, function, and information processing Matter and energy in organisms and ecosystems Interdependent relationships in ecosystems Growth, development, and reproduction of organisms Natural selection and adaptations Engineering design in life science	2, 3	34–40
	Earth and space science	Space systems History of the Earth Earth's systems Weather and climate Human impacts Engineering design in Earth and space science	2, 3	27–33
11	Physical science	Structure and properties of matter Chemical reactions Forces and interactions Energy Waves and electromagnetic radiation Engineering design in physical science	2, 3	27–33
	Life science	Structure and function Matter and energy in organisms and ecosystems Interdependent relationships in ecosystems Inheritance and variation of traits Natural selection and evolution Engineering design in life science	2, 3	34–40

Grade	Claim	Target	Depth of knowledge	% of items
	Earth and space science	Space systems History of the Earth Earth's systems Weather and climate Human sustainability Engineering design in Earth and space science	2, 3	27–33

Appendix B: Item Statistics Flagging Criteria

Table B-1. Item Statistics Flagging Criteria

Statistic	Criteria
Omit	Omit correlation > .10 Omit percentage > .05
Differential item functioning	Gender R^2 change > 0.035 Race R^2 change > 0.035 Ethnicity R^2 change > 0.035 English learner R^2 change > 0.035
Item total correlation	Item total correlation ≤ .25
p value	*p* value < .20 *p* value > .90
Distractors for selecting-key items	Correlation of distractors > −0.05 Percentage of selecting distractor > Percentage of selecting keyed response

KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

Appendix C: Subjects Performance Level Descriptors (PLDs)

Grades 3–8 and 10 English language arts			
Level 1	Level 2	Level 3	Level 4
A student at level 1 shows a limited ability to understand and use the skills and knowledge needed for postsecondary readiness.	A student at level 2 shows a basic ability to understand and use the skills and knowledge needed for postsecondary readiness.	A student at level 3 shows an effective ability to understand and use the skills and knowledge needed for postsecondary readiness.	A student at level 4 shows an excellent ability to understand and use the skills and knowledge needed for college and career readiness.
Level 1 scores are difficult to interpret. They range from no correct answers to scores that miss level 2 by one point. There are a number of possible reasons a student's performance resulted in a level 1 score. However, students whose scores fall into level 1 typically have difficulty reading, analyzing, and understanding grade-level texts; editing a text to use appropriate general language, grammar, and punctuation using writing strategies appropriate for different types of texts; and structuring a text to support a purpose or opinion. Parents/guardians and teachers are encouraged to examine other academic information and discuss possible reasons that a student's score is in level 1.	Students who score at level 2 can typically • read and understand readily accessible grade-level texts, • identify central ideas and clear details, • determine meanings of common words and straightforward figurative language, • identify text features and structures that organize a text, • identify relationships between parts of a text, • revise or edit a text to use appropriate general language, grammar, and punctuation, • use writing strategies appropriate for different types of text, and • structure a text to support a purpose or opinion.	Students who score at level 3 can typically • 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 and punctuation, • organize a text using sequence and logic, • determine if information is relevant, and • use strategies to elaborate on ideas and structure texts.	Students who score at level 4 can typically • read and understand very complex grade-level texts, • summarize and analyze themes, point of view, and purpose, • use implied and clear details to support or refute an inference or conclusion, • interpret and analyze literary devices and word choice and their impact on meaning and tone, • revise and edit a text to use challenging vocabulary and correct grammar and punctuation, • organize details or elaborate on ideas for a purpose, and • show understanding of appropriate text structure.

Grade-3 Mathematics			
Level 1	Level 2	Level 3	Level 4
A student at level 1 shows a limited ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 2 shows a basic ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 3 shows an effective ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 4 shows an excellent ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.
Students who score at level 1 can typically • use multiplication and division to solve problems involving equal groups of objects, • add and subtract within 100, • round two-digit whole numbers to the nearest 10, • identify fractions on a number line, • tell and write time to five-minute intervals, • represent data sets using picture graphs, bar graphs, and line plots, • recognize area as an attribute of plane figures, and • determine perimeter of polygons.	Students who score at level 2 can typically • represent and interpret multiplication problems using models, • recall products within the 10-by-10 multiplication table, • perform operations to solve one- and two-step word problems, • round whole numbers to the nearest 100, • identify equivalent fractions, • tell and write time to the nearest minute, • solve one-step problems involving data represented in bar graphs, • determine area of rectangles using unit squares, and • determine an unknown side length of a polygon using the perimeter and known side lengths.	Students who score at level 3 can typically • fluently multiply and divide within 100, • solve division problems with an unknown factor, • represent word problems using equations, • add and subtract within 1,000 using a variety of strategies, • create equivalent fractions and compare two fractions, • solve addition and subtraction problems involving intervals of time in minutes, • solve two-step problems involving data represented in bar graphs, • determine area of rectangles and rectilinear figures (figures made of straight lines), and • identify rectangles based on their perimeter and area.	Students who score at level 4 can typically • use multiplication and division to solve problems involving a two-digit factor, • use properties of operations to multiply within 100 using a two-digit factor, • approximate fractions on a number line without partitioning, • explain why two fractions are equivalent, • solve addition and subtraction problems involving intervals of time in hours and minutes, and • solve real-world problems involving perimeter and area of rectangles and area of rectilinear figures (figures made of straight lines).

KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

Grade-4 Mathematics			
Level 1	Level 2	Level 3	Level 4
A student at level 1 shows a limited ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 2 shows a basic ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 3 shows an effective ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 4 shows an excellent ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.
Students who score at level 1 can typically • represent and solve one-step word problems, • read, write, and round multidigit whole numbers in various forms, • compare multidigit whole numbers written in the same form, • perform operations with one- and two-digit whole numbers, • recognize fraction comparisons must refer to the same whole, • identify fractions using visual models, • identify information presented in line plots, bar graphs, and pictographs, and • draw lines and types of angles.	Students who score at level 2 can typically • represent and solve two-step word problems, • read, write, and round multidigit whole numbers in various forms, • compare multidigit whole numbers written in the same form, • identify equivalent fractions using visual models, • represent and solve problems involving addition and subtraction of fractions, • write fractions with denominators 10 or 100 as decimals, • determine perimeter and area of rectangles, • represent data sets using line plots, bar graphs, and pictographs, and • identify lines and angles in two-dimensional figures.	Students who score at level 3 can typically • represent and solve multistep word problems, • read, write, and round multidigit whole numbers in various forms, • compare multidigit whole numbers written in different forms, • perform operations with multidigit whole numbers, • create equivalent fractions using visual models, • add and subtract fractions and mixed numbers, • locate decimal numbers on a number line, • know and apply area and perimeter formulas to determine the missing side length of a rectangle, • solve problems using information presented in line plots, bar graphs, and pictographs, and • classify two-dimensional figures.	Students who score at level 4 can typically • use mental computation and estimation strategies to determine if answers are reasonable, • explain multiplication and division using equations, models, place-value understanding, and properties of operations, • compare fractions and justify the comparisons, • solve multistep problems involving multiplication of a fraction by a whole number, • compare decimals and justify the comparisons, • solve real-world problems involving perimeter and area of rectangles, • interpret information presented in line plots, bar graphs, and pictographs, and • categorize triangles based on angles and sides.

Grade-5 Mathematics			
Level 1	Level 2	Level 3	Level 4
A student at level 1 shows a limited ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 2 shows a basic ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 3 shows an effective ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 4 shows an excellent ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.
Students who score at level 1 can typically • read, write, and round decimals in various forms, • multiply and divide multidigit whole numbers, • perform operations with decimals and whole numbers, • add and subtract fractions, • multiply whole numbers by fractions, • identify information presented in line plots, bar graphs, and pictographs, • determine volume of right rectangular prisms using unit cubes, and • graph whole-number coordinate points on a coordinate grid.	Students who score at level 2 can typically • compare decimals written in the same form, • multiply and divide multidigit whole numbers, • perform operations with decimals and whole numbers, • add fractions and mixed numbers, • multiply whole numbers by fractions and mixed numbers, • represent data sets using line plots, bar graphs, and pictographs, • represent volume of right rectangular prisms as the prisms' edge lengths, and • graph whole-number coordinate points on a coordinate grid.	Students who score at level 3 can typically • represent powers of 10 using exponents, • multiply and divide multidigit whole numbers, • perform operations with decimals, • add and subtract fractions and mixed numbers, • multiply fractions and mixed numbers, and divide unit fractions by whole numbers, • solve problems using information presented in line plots, bar graphs, and pictographs, • determine volume of right rectangular prisms, and • interpret whole-number coordinate points on a coordinate grid.	Students who score at level 4 can typically • solve real-world and mathematical problems involving powers of 10, • perform operations with decimals and justify the calculations, • add and subtract fractions and mixed numbers to solve word problems, • multiply fractions and mixed numbers, and divide unit fractions and whole numbers to solve word problems, • interpret information presented in line plots, bar graphs, and pictographs, • compare the volumes of two rectangular prisms, and • graph and interpret coordinate points of fractions on a coordinate grid.

Grade-6 Mathematics			
Level 1	Level 2	Level 3	Level 4
A student at level 1 shows a limited ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 2 shows a basic ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 3 shows an effective ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 4 shows an excellent ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.
Students who score at level 1 can typically • describe ratio relationships between two quantities, • divide fractions by whole numbers, • add, subtract, and multiply whole numbers and decimals, • locate integers on a number line, • write and evaluate numerical expressions, • use substitution to determine solutions to equations, • identify a table of values that represents a relationship between two variables, • determine area of right triangles and volume of rectangular prisms, and • summarize data using dot plots and histograms.	Students who score at level 2 can typically • use ratio reasoning to determine unit rate, • divide whole numbers by fractions, • add, subtract, and multiply multidigit decimals, • graph ordered pairs of integers on a coordinate plane, • write and evaluate numerical expressions with exponents, • write and solve algebraic equations, • use graphs and tables to represent a relationship between two variables, • determine area of special quadrilaterals and triangles and volume of rectangular prisms, and • summarize data using stem-and-leaf plots.	Students who score at level 3 can typically • determine and use unit rates to solve multistep problems, • divide fractions and mixed numbers, • perform operations with multidigit decimals, • graph ordered pairs of rational numbers on a coordinate plane, • write and evaluate numerical and algebraic expressions, • use graphs, tables, and context to analyze linear relationships, • determine area of polygons, surface area of nets, and volume of rectangular prisms, and • summarize data using box plots.	Students who score at level 4 can typically • explain ratio relationships between two quantities, • solve real-world problems involving division of fractions and interpret solutions, • use properties to show why expressions are equivalent, • use graphs, tables, and context to analyze linear relationships, • determine surface area and volume of figures composed of rectangular prisms, and • justify the reasonableness of the center and spread of a data set.

Grade-7 Mathematics			
Level 1	Level 2	Level 3	Level 4
A student at level 1 shows a limited ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 2 shows a basic ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 3 shows an effective ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 4 shows an excellent ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.
Students who score at level 1 can typically • identify proportional relationships, • perform operations with rational numbers using a number line, • add and subtract linear expressions, • solve equations with integers, • determine area of triangles and rectangles and volume of cubes, • use the mean to compare two populations, and • determine probability of simple events.	Students who score at level 2 can typically • represent proportional relationships using equations, • factor and expand linear expressions, • write and solve equations with rational numbers, • determine the scale factor between a geometric figure and its scale drawing, • determine circumference and area of circles and volume of cylinders, • use data from a random sample to make inferences about a population, and • determine probability of chance events using data.	Students who score at level 3 can typically • analyze proportional relationships presented in a variety of ways, • use the four operations to solve real-world problems involving rational numbers, • factor and expand linear expressions, • solve and graph inequalities in one variable, • create scale drawings of geometric figures, • solve problems involving volume and surface area of three-dimensional figures, • use data from a random sample to make inferences about two populations, and • determine probability of compound events.	Students who score at level 4 can typically • solve real-world problems involving proportional relationships, • solve and interpret real-world problems involving rational numbers, • interpret solution sets to inequalities in one variable, • describe two-dimensional figures made from slicing three-dimensional figures, • solve real-world problems involving volume and surface area of three-dimensional figures, • use multiple samples to estimate and make predictions about a population, and • explain possible differences between theoretical probability and experimental results.

KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

Grade-8 Mathematics			
Level 1	Level 2	Level 3	Level 4
A student at level 1 shows a limited ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 2 shows a basic ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 3 shows an effective ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 4 shows an excellent ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.
Students who score at level 1 can typically • classify numbers as rational or irrational, • graph proportional relationships, • solve one- and two-step linear equations and inequalities in one variable, • determine whether relations, presented in various formats, are functions, • represent linear relationships using graphs or tables and calculate rate of change, • measure angles and classify pairs of angles as supplementary or complementary, • identify the hypotenuse and legs of a right triangle, • identify key dimensions of pyramids, cones, and spheres, and • construct scatter plots.	Students who score at level 2 can typically • convert between fractions and terminating decimals, • calculate slope of a line using two coordinate points, • solve multistep linear equations in one variable, • produce input-output pairs for functions, • construct linear functions, • solve problems involving unknown angle measurements, • determine whether a triangle is a right triangle, • recognize formulas for volume and surface area of pyramids, cones, and spheres, and • informally fit a line to data of linear association.	Students who score at level 3 can typically • convert between fractions and repeating decimals, • convert between standard form and scientific notation, • solve multistep linear equations and inequalities in one variable, • classify functions as linear or nonlinear, • determine rate of change and initial value of linear functions, • compare two linear functions, • determine an unknown side length of a right triangle, • apply volume and surface area formulas for pyramids, cones, and spheres, and • interpret scatter plots and describe patterns of association.	Students who score at level 4 can typically • approximate irrational numbers, • use scientific notation to estimate and compare quantities, • describe the relationship between proportional and non-proportional linear relationships, • write and solve multistep linear inequalities in one variable, • give examples of functions that are not linear, • analyze graphs of nonlinear functions, • generalize relationships of angles when parallel lines are cut by a transversal, • apply the Pythagorean theorem to determine the distance between two points, and • use scatter plots, trend lines, and associations to make predictions in real-world situations.

Grade-10 Mathematics			
Level 1	Level 2	Level 3	Level 4
A student at level 1 shows a limited ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 2 shows a basic ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 3 shows an effective ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.	A student at level 4 shows an excellent ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.
Students who score at level 1 can typically • factor quadratic expressions, • add, subtract, and multiply single-variable binomials, • solve linear equations in one variable, • graph systems of two linear equations and estimate solutions, • graph linear functions and interpret key features of the functions, • identify transformations of figures, • identify components of triangles and parallelograms to construct arguments related to geometric theorems, and • describe data in terms of center and spread.	Students who score at level 2 can typically • write quadratic expressions in equivalent forms, • add, subtract, and multiply single-variable trinomials, • solve linear inequalities in one variable, • recognize the number of solutions for systems of two linear equations, • graph quadratic functions and interpret key features of the functions, • identify sequences of transformations on figures, • identify properties to construct arguments related to geometric theorems, and • describe data sets in terms of shape, center, and spread.	Students who score at level 3 can typically • determine and use the zeros of a factored quadratic expression to solve problems, • add, subtract, and multiply multivariable polynomials (expressions that include variables and exponents), • solve quadratic, absolute value, and simple rational equations in one variable, • solve systems of two linear equations, • graph absolute value functions and interpret key features of the functions, • describe the effects of transformations on figures, • construct arguments related to geometric theorems and complete proofs, and • use appropriate statistics to compare sets of data.	Students who score at level 4 can typically • identify appropriate equivalent forms of quadratic expressions to reveal different properties, • add, subtract, and multiply multivariable polynomials (expressions that include variables and exponents), • solve literal equations for a specified quantity, • compare properties of two different types of functions, • recognize transformations as functions, • explain why two figures are similar or congruent in relation to a sequence of transformations, • identify errors in proofs, and • interpret data and explain why a data value is an outlier.

Grade-5 Science			
Level 1	Level 2	Level 3	Level 4
A student at level 1 shows a limited ability to understand and use the science skills and knowledge needed for postsecondary readiness.	A student at level 2 shows a basic ability to understand and use the science skills and knowledge needed for postsecondary readiness.	A student at level 3 shows an effective ability to understand and use the science skills and knowledge needed for postsecondary readiness.	A student at level 4 shows an excellent ability to understand and use the science skills and knowledge needed for postsecondary readiness.
Level 1 scores are difficult to interpret. They range from no correct answers to scores that miss level 2 by one point. There are a number of possible reasons a student's performance resulted in a level 1 score; however, students whose scores fall into level 1 typically have difficulty understanding, explaining, and analyzing complex grade-level science concepts and practices.	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, and • 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 1	Level 2	Level 3	Level 4
A student at level 1 shows a limited ability to understand and use the science skills and knowledge needed for postsecondary readiness.	A student at level 2 shows a basic ability to understand and use the science skills and knowledge needed for postsecondary readiness.	A student at level 3 shows an effective ability to understand and use the science skills and knowledge needed for postsecondary readiness.	A student at level 4 shows an excellent ability to understand and use the science skills and knowledge needed for postsecondary readiness.
level 1 scores are difficult to interpret. They range from no correct answers to scores that miss level 2 by one point. There are a number of possible reasons a student's performance resulted in a level 1 score; however, students whose scores fall into level 1 typically have difficulty understanding, explaining, and analyzing complex grade-level science concepts and practices.	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			
Level 1	Level 2	Level 3	Level 4
A student at level 1 shows a limited ability to understand and use the science skills and knowledge needed for postsecondary readiness.	A student at level 2 shows a basic ability to understand and use the science skills and knowledge needed for postsecondary readiness.	A student at level 3 shows an effective ability to understand and use the science skills and knowledge needed for postsecondary readiness.	A student at level 4 shows an excellent ability to understand and use the science skills and knowledge needed for postsecondary readiness.
Level 1 scores are difficult to interpret. They range from no correct answers to scores that miss level 2 by one point. There are a number of possible reasons a student's performance resulted in a level 1 score; however, students whose scores fall into level 1 typically have difficulty understanding, explaining, and analyzing complex grade-level science concepts and practices.	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.

Appendix D: Subscore Reliability

Table D-1. English Language Arts Subscore, Reliability, Classification Consistency, and Accuracy by Grade

Grade	Subscore name	Reliability	Consistency	Accuracy
3	Overall Reading	.70	.44	.76
3	Reading: Key Ideas & Details	.64	.34	.71
3	Reading: Craft, Structure, & Language in Reading	.64	.41	.75
3	Overall Writing	.49	.33	.70
3	Writing: Text Types and Purposes	.54	.30	.69
3	Writing: Language in Writing	.60	.35	.72
4	Overall Reading	.70	.41	.74
4	Reading: Key Ideas & Details	.65	.38	.72
4	Reading: Craft, Structure, & Language in Reading	.63	.33	.69
4	Overall Writing	.46	.31	.67
4	Writing: Text Types and Purposes	.55	.31	.70
4	Writing: Language in Writing	.54	.25	.66
5	Overall Reading	.70	.42	.75
5	Reading: Key Ideas & Details	.67	.38	.72
5	Reading: Craft, Structure, & Language in Reading	.60	.37	.74
5	Overall Writing	.61	.35	.69
5	Writing: Text Types and Purposes	.55	.30	.64
5	Writing: Language in Writing	.63	.39	.75
6	Overall Reading	.69	.39	.76
6	Reading: Key Ideas & Details	.67	.38	.77
6	Reading: Craft, Structure, & Language in Reading	.58	.29	.72
6	Overall Writing	.59	.30	.67
6	Writing: Text Types and Purposes	.55	.33	.71
6	Writing: Language in Writing	.57	.36	.69

Grade	Subscore name	Reliability	Consistency	Accuracy
7	Overall Reading	.67	.38	.79
7	Reading: Key Ideas & Details	.61	.33	.76
7	Reading: Craft, Structure, & Language in Reading	.61	.35	.74
7	Overall Writing	.61	.30	.68
7	Writing: Text Types and Purposes	.61	.34	.68
7	Writing: Language in Writing	.51	.25	.63
8	Overall Reading	.66	.36	.82
8	Reading: Key Ideas & Details	.58	.29	.80
8	Reading: Craft, Structure, & Language in Reading	.61	.36	.82
8	Overall Writing	.65	.40	.81
8	Writing: Text Types and Purposes	.62	.41	.81
8	Writing: Language in Writing	.61	.41	.75
10	Overall Reading	.68	.36	.78
10	Reading: Key Ideas & Details	.65	.35	.78
10	Reading: Craft, Structure, & Language in Reading	.58	.33	.79
10	Overall Writing	.62	.36	.79
10	Writing: Text Types and Purposes	.60	.37	.80
10	Writing: Language in Writing	.55	.48	.88

Table D-2. Mathematics Subscore, Reliability, Classification Consistency, and Accuracy by Grade

Grade	Subscore name	Reliability	Consistency	Accuracy
3	Skills and Concepts	.80	.49	.76
3	Operations and Algebraic Thinking	.71	.41	.70
3	Geometry	.69	.35	.71
3	Number and Operations With Fractions	.62	.36	.71
3	Measurement and Data	.70	.37	.69
3	Strategic Thinking and Reasoning	.55	.29	.63
4	Skills and Concepts	.79	.51	.81
4	Operations and Algebraic Thinking	.64	.33	.72
4	Number and Operations in Base Ten	.65	.35	.70
4	Number and Operations With Fractions	.73	.45	.77
4	Measurement and Data	.53	.25	.64
4	Strategic Thinking and Reasoning	.55	.19	.58
5	Skills and Concepts	.77	.53	.84
5	Number and Operations in Base Ten	.67	.43	.77
5	Number and Operations With Fractions	.66	.37	.77
5	Measurement and Data	.64	.38	.75
5	Strategic Thinking and Reasoning	.54	.27	.70
6	Skills and Concepts	.77	.50	.82
6	Geometry	.59	.26	.73
6	Statistics and Probability	.59	.34	.75
6	Ratios and Proportional Relationships	.60	.36	.75
6	The Number System	.66	.39	.77
6	Expressions and Equations	.67	.39	.78
6	Strategic Thinking and Reasoning	.58	.29	.77

7	Skills and Concepts	.74	.50	.84
7	Geometry	.59	.29	.78
7	Statistics and Probability	.61	.33	.79
7	Ratios and Proportional Relationships	.54	.26	.75
7	The Number System	.65	.39	.79
7	Expressions and Equations	.66	.39	.77
7	Strategic Thinking and Reasoning	.52	.26	.71
8	Skills and Concepts	.71	.49	.87
8	Geometry	.59	.32	.78
8	Expressions and Equations	.64	.38	.81
8	Functions	.62	.39	.80
8	Strategic Thinking and Reasoning	.57	.30	.74
10	Skills and Concepts	.72	.54	.88
10	Geometry	.66	.41	.82
10	Statistics and Probability	.59	.37	.74
10	Algebra	.64	.42	.84
10	Functions	.51	.32	.80
10	Strategic Thinking and Reasoning	.50	.24	.74

Table D-3. Science Subscore, Reliability, Classification Consistency, and Accuracy by Grade

Grade	Subscore name	Reliability	Consistency	Accuracy
5	Physical and chemical sciences	.63	.35	.68
5	Life sciences	.59	.29	.64
5	Earth and space sciences	.65	.36	.70
8	Physical and chemical sciences	.56	.30	.81
8	Life sciences	.59	.35	.77
8	Earth and space sciences	.53	.28	.77
11	Physical and chemical sciences	.62	.38	.76
11	Life sciences	.65	.42	.76
11	Earth and space sciences	.58	.33	.75

Appendix E: School Board of Education District Demographic Distribution

Table E-1. Number of Students Enrolled and Their Demographic Distribution by State Board of Education District

District	*N*	%													
		Gender		Race						Hispanic		SWD		EL	
		Female	Male	NA	Asian	Black	NHPI	Other	White	No	Yes	No	Yes	No	Yes
1	49895	49	51	3	3	13	0	8	73	73	27	77	23	84	16
2	52373	49	51	1	6	7	0	6	79	84	16	90	10	91	9
3	60957	49	51	1	6	6	0	6	81	86	14	89	11	93	7
4	31730	49	51	2	1	7	0	10	79	85	15	84	16	95	5
5	32021	49	51	7	1	2	0	4	86	58	42	86	14	76	24
6	32105	49	51	2	1	5	1	7	84	89	11	83	17	97	3
7	66419	49	51	2	3	11	0	8	75	76	24	83	17	89	11
8	39201	49	51	2	5	16	0	10	67	71	29	85	15	84	16
9	36326	49	51	2	1	2	0	7	88	92	8	83	17	97	3
10	63282	49	51	2	3	11	0	8	76	77	23	84	16	89	11

Note. NA = Native American; NHPI = Native Hawaiian and Pacific Islander; EL = English learner; SWD = student with disability.

Appendix F: Sample KAP Reports

Figure F-1. Sample KAP Student Report: Mathematics

STUDENT REPORT: Smith, Sean
GRADE: 10 Mathematics / STATE ID: 123456789
SCHOOL: Meadowlark School
DISTRICT: Sunflower District / #D1001

2021–2022
KAP Kansas Assessment Program

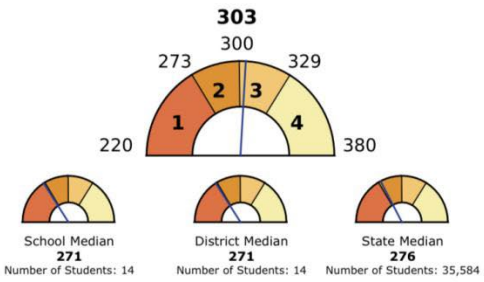
The KAP assessments measure students' understanding of the Kansas Standards at each grade. The math assessment asks students to answer computation questions and questions about data presented in word problems, equations, graphs, tables, and diagrams. Students may show what they know about mathematics by selecting or providing the right answer, sorting or ordering items, creating graphs, and labeling pictures.

Mathematics Score: Level 3

1 2 3 4

Your student's performance level from last year is not available.

When interpreting KAP results, please take into consideration how the conditions for learning, which may have been disrupted by the pandemic, may influence performance.



Students who score at this level can typically

- determine and use the zeros of a factored quadratic expression to solve problems
- add, subtract, and multiply multivariable polynomials (expressions that include variables and exponents)
- solve quadratic, absolute value, and simple rational equations in one variable
- solve systems of two linear equations
- graph absolute value functions and interpret key features of the functions
- describe the effects of transformations on figures
- construct arguments related to geometric theorems and complete proofs
- use appropriate statistics to compare sets of data

Standard error of measurement for this report:
Student — 6.9 | School — 6.6 | District — 6.6 | 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).

Performance Level Descriptions

Level 1: A student at Level 1 shows a limited ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.

Level 2: A student at Level 2 shows a basic ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.

Level 3: A student at Level 3 shows an effective ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.

Level 4: A student at Level 4 shows an excellent ability to understand and use the mathematics skills and knowledge needed for postsecondary readiness.

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

KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

STUDENT REPORT | STUDENT: Smith, Sean | STATE ID: 123456789 | GRADE: 10 Mathematics

Your Student's Performance

(+) Exceeds (=) Meets (−) Below (×) Insufficient Data

SKILLS AND CONCEPTS

(=) **In this area, your students typically performed as well as students who received the minimum Level 3 score.** These questions require students to apply mathematical skills and concepts and interpret and carry out mathematical procedures with precision and fluency.

Algebra

(−) **In this area, your students typically performed below students who received the minimum Level 3 score.** These questions require students to solve complex equations; construct, interpret, and graph equations that model data and represent relationships; and use equations to solve real-world problems.

Functions

(=) **In this area, your students typically performed as well as students who received the minimum Level 3 score.** These questions require students to interpret, compare, and build functions to model real-world relationships.

Geometry

(+) **In this area, your students typically performed better than students who received the minimum Level 3 score.** These questions require students to describe the features of geometric figures, compare figures, apply geometric theorems, and solve real-world problems by applying formulas to figures.

Statistics and Probability

(−) **In this area, your students typically performed below students who received the minimum Level 3 score.** These questions require students to compare and draw inferences from data sets and to calculate probability of simple and compound events.

STRATEGIC THINKING AND REASONING

(−) **In this area, your students typically performed below students who received the minimum Level 3 score.** These questions require students to solve complex problems using problem-solving strategies and mathematical tools; explain their reasoning, defend their answers, and critique the reasoning of others; and analyze complex, real-world situations to construct and use mathematical models to solve problems, and to interpret results in the context of a situation.

Additional Resources

To learn more about the Kansas Assessment Program and these score reports, visit the "For Families" page on **ksassessments.org**. For information on the Kansas Standards, visit **ksde.org**.

Prediction on ACT scores is not available for mathematics grade 10 in 2022.

Quantile® Measure

Your student's score: **1205Q**

The Quantile measure provides a score that describes your child's level of mathematical ability and the difficulty of a skill or concept as it relates to other mathematical skills and concepts your child is learning. The score shows your child's readiness for instruction regarding a particular mathematical skill or concept.

Kansans CAN

Figure F-2. Sample KAP Student Report: Science

STUDENT REPORT: Smith, Sean
GRADE: 8 Science / STATE ID: 123456789
SCHOOL: Meadowlark School
DISTRICT: Sunflower District / #D1001

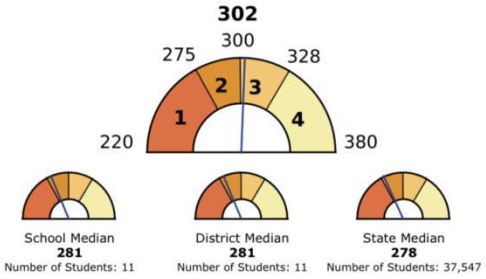
2021—2022
KAP Kansas Assessment Program

The KAP assessments measure students' understanding of the Kansas Standards at each grade. The science assessment asks students to answer questions about data presented in narratives, equations, graphs, tables, and diagrams. Students show what they know about science by selecting or providing the right answer; sorting, ordering, or matching items; and labeling pictures.

Science Score: Level 3

1 2 **3** 4

When interpreting KAP results, please take into consideration how the conditions for learning, which may have been disrupted by the pandemic, may influence performance.



Students who score at this level 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
- develop a model to optimize an engineering design through repeated testing

Standard error of measurement for this report:
Student — 10.8 | School — 11.3 | District — 11.3 | 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).

Performance Level Descriptions

Level 1: A student at Level 1 shows a limited ability to understand and use the science skills and knowledge needed for postsecondary readiness.

Level 2: A student at Level 2 shows a basic ability to understand and use the science skills and knowledge needed for postsecondary readiness.

Level 3: A student at Level 3 shows an effective ability to understand and use the science skills and knowledge needed for postsecondary readiness.

Level 4: A student at Level 4 shows an excellent ability to understand and use the science skills and knowledge needed for postsecondary readiness.

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

KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

STUDENT REPORT

STUDENT: Smith, Sean
STATE ID: 123456789

GRADE: 8 Science

Your Student's Performance

Exceeds | Meets | Below | Insufficient Data

PHYSICAL AND CHEMICAL SCIENCES

In this area, your students typically performed below students who received the minimum Level 3 score. These 3-dimensional questions about phenomena require students to understand and apply (1) practices in science and engineering (ex. Analyzing and Interpreting Data), (2) their core ideas (ex. Chemical Reactions), and (3) concepts that crosscut science disciplines (ex. Stability and Change).

LIFE SCIENCES

In this area, your students typically performed as well as students who received the minimum Level 3 score. These 3-dimensional questions about phenomena require students to understand and apply (1) practices in science and engineering (ex. Engaging in Argument from Evidence), (2) their core ideas (ex. Ecosystem Relationships), and (3) concepts that crosscut science disciplines (ex. Energy and Matter).

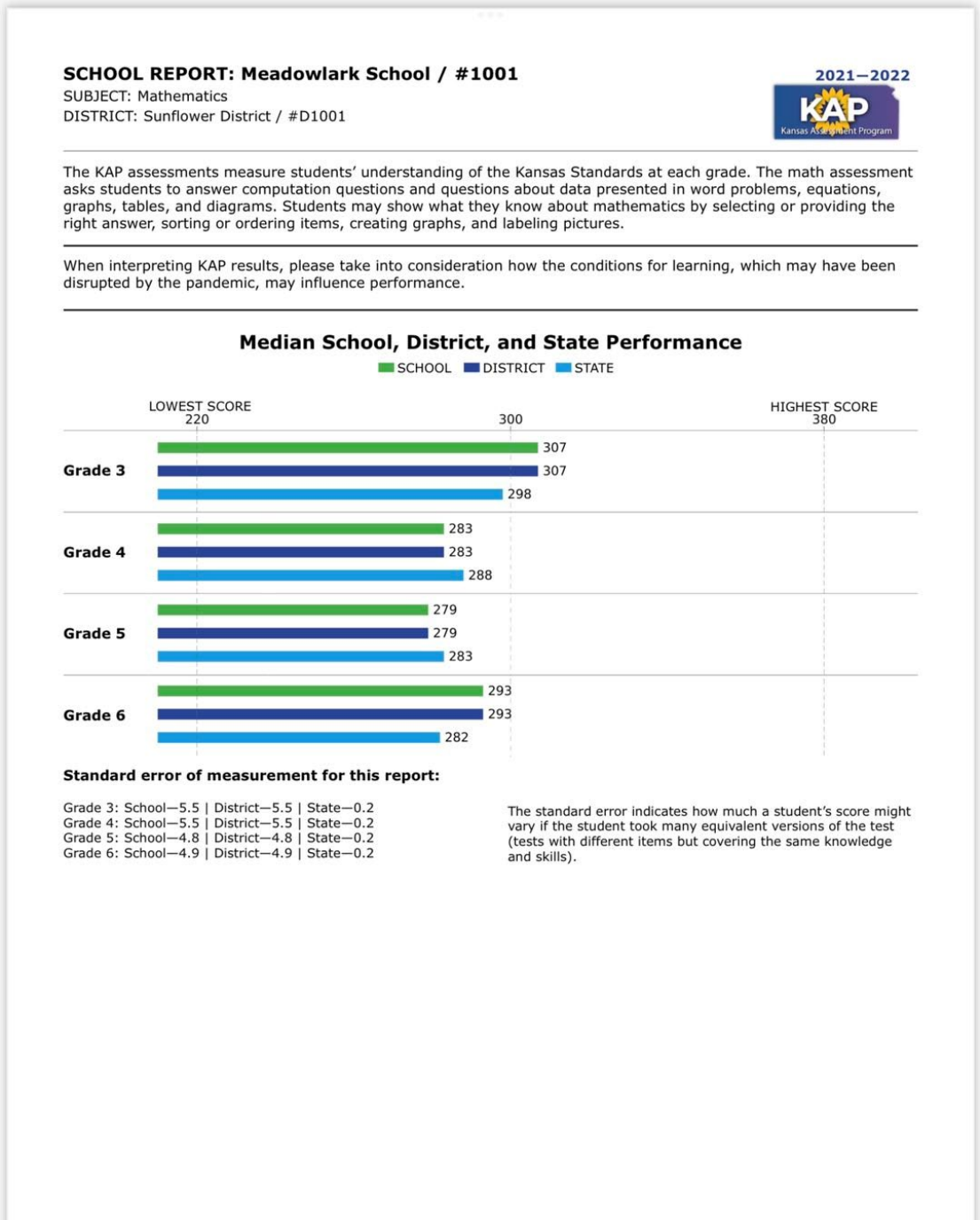
EARTH AND SPACE SCIENCES

In this area, your students typically performed as well as students who received the minimum Level 3 score. These 3-dimensional questions about phenomena require students to understand and apply (1) practices in science and engineering (ex. Developing and Using Models), (2) their core ideas (ex. Earth Systems), and (3) concepts that crosscut science disciplines (ex. Systems and System Models).

Additional Resources
To learn more about the Kansas Assessment Program and these score reports, visit the "For Families" page on **ksassessments.org**.
For information on the Kansas Standards, visit **ksde.org**.

Kansans CAN

Figure F-3. Sample KAP School Report



SCHOOL REPORT: Meadowlark School / #1001
SUBJECT: Mathematics
DISTRICT: Sunflower District / #D1001

2021—2022

The KAP assessments measure students' understanding of the Kansas Standards at each grade. The math assessment asks students to answer computation questions and questions about data presented in word problems, equations, graphs, tables, and diagrams. Students may show what they know about mathematics by selecting or providing the right answer, sorting or ordering items, creating graphs, and labeling pictures.

When interpreting KAP results, please take into consideration how the conditions for learning, which may have been disrupted by the pandemic, may influence performance.

Median School, District, and State Performance

Standard error of measurement for this report:

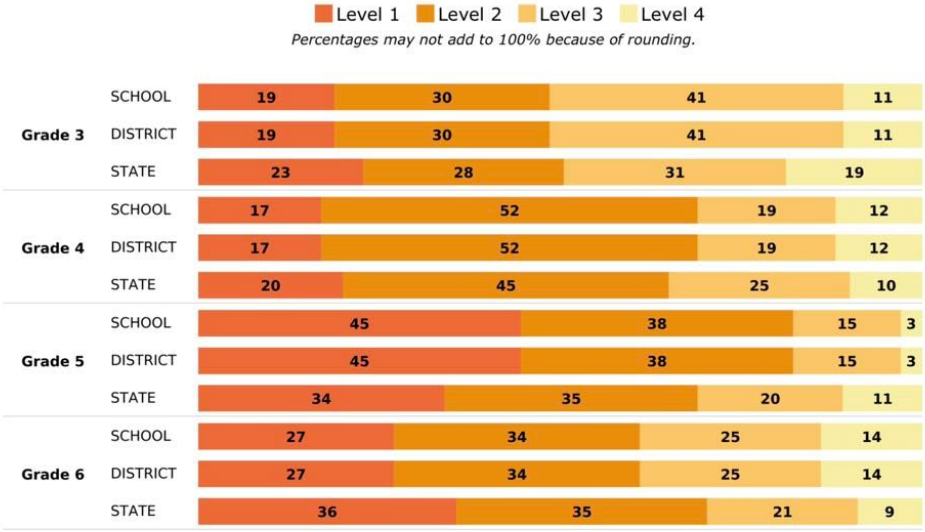
Grade 3: School—5.5 | District—5.5 | State—0.2
Grade 4: School—5.5 | District—5.5 | State—0.2
Grade 5: School—4.8 | District—4.8 | State—0.2
Grade 6: School—4.9 | District—4.9 | State—0.2

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).

KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

SCHOOL REPORT SCHOOL: MEADOWLARK SCHOOL

Percentage of Students in Each Performance Level, by Grade



KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

SCHOOL REPORT SCHOOL: MEADOWLARK SCHOOL

Your School's Performance

+ Exceeds = Meets − Below ⊗ Insufficient Data

Grade	3	4	5	6
SKILLS AND CONCEPTS	=	−	−	−
Operations and Algebraic Thinking	=	−		
Number and Operations in Base Ten		−	−	
Number and Operations with Fractions	=	−	−	
Measurement and Data	=	−	−	
Ratios and Proportional Relationships				=
The Number System				−
Expressions and Equations				−
Geometry	−			−
Statistics and Probability				−
STRATEGIC THINKING AND REASONING	−	−	−	−

SKILLS AND CONCEPTS
These questions require students to apply mathematical skills and concepts and interpret and carry out mathematical procedures with precision and fluency.

Operations and Algebraic Thinking
These questions require students to represent and solve problems with addition, subtraction, multiplication, and division; perform these operations with multidigit numbers; and explain patterns.

Number and Operations in Base Ten
These questions require students to demonstrate their understanding of place value by solving problems with multidigit numbers and decimals.

Number and Operations with Fractions
These questions require students to demonstrate their understanding that fractions represent parts of a whole, recognize that fractions can be written as decimals, and solve problems with fractions by applying their knowledge about working with whole numbers and decimals.

Measurement and Data
These questions require students to calculate time, volume, perimeter, area, and mass; measure angle size; convert measurements within a measurement system; represent and interpret measurement data; and use measurement skills to solve real-world problems.

Ratios and Proportional Relationships
These questions require students to use ratio reasoning and analyze proportional relationships to solve real-world and mathematical problems.

The Number System
These questions require students to divide fractions, find common factors and multiples, and perform operations with rational numbers.

Expressions and Equations
These questions require students to solve equations that have variables and exponents, analyze relationships between dependent and independent variables and between proportional relationships, and use equations to model relationships and solve real-world problems.

KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

Geometry
These questions require students to describe the features of geometric figures, compare figures, apply geometric theorems, and solve real-world problems by applying formulas to figures.

Statistics and Probability
These questions require students to compare and draw inferences from data sets and to calculate probability of simple and compound events.

STRATEGIC THINKING AND REASONING
These questions require students to solve complex problems using problem-solving strategies and mathematical tools; explain their reasoning, defend their answers, and critique the reasoning of others; and analyze complex, real-world situations to construct and use mathematical models to solve problems, and to interpret results in the context of a situation.

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.

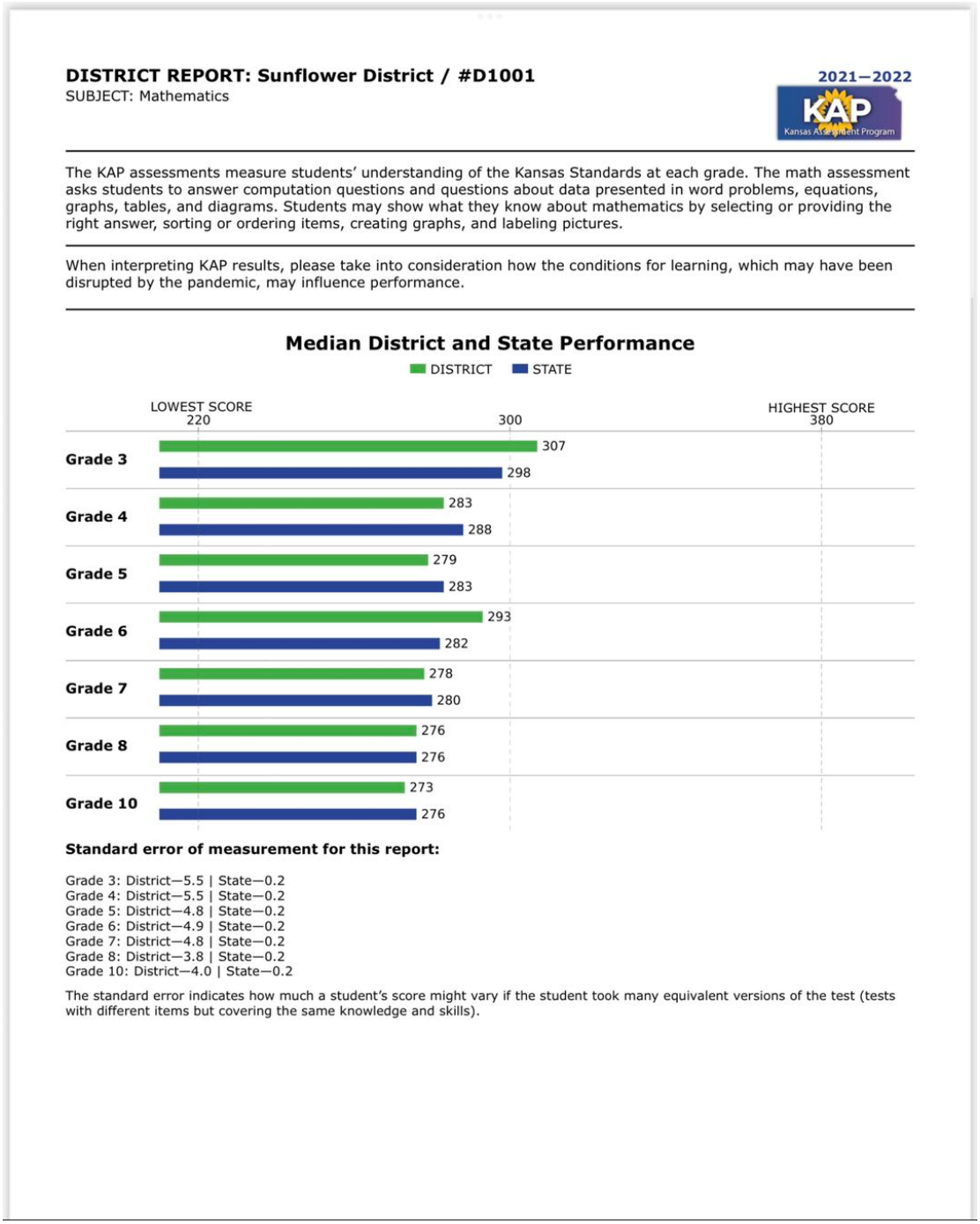
Insufficient Data
In this area, your students did not answer enough questions for accurate reporting.

Additional Resources
To learn more about the Kansas Assessment Program and these score reports, visit the "For Families" page on **ksassessments.org**.
For information on the Kansas Standards, visit **ksde.org**.

Kansans CAN

KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

Figure F-4. Sample KAP District Report



DISTRICT REPORT: Sunflower District / #D1001
SUBJECT: Mathematics

2021–2022
KAP Kansas Assessment Program

The KAP assessments measure students' understanding of the Kansas Standards at each grade. The math assessment asks students to answer computation questions and questions about data presented in word problems, equations, graphs, tables, and diagrams. Students may show what they know about mathematics by selecting or providing the right answer, sorting or ordering items, creating graphs, and labeling pictures.

When interpreting KAP results, please take into consideration how the conditions for learning, which may have been disrupted by the pandemic, may influence performance.

Median District and State Performance

DISTRICT STATE

LOWEST SCORE 220 | 300 | HIGHEST SCORE 380

	DISTRICT	STATE
Grade 3	307	298
Grade 4	283	288
Grade 5	279	283
Grade 6	293	282
Grade 7	278	280
Grade 8	276	276
Grade 10	273	276

Standard error of measurement for this report:

Grade 3: District—5.5 | State—0.2
Grade 4: District—5.5 | State—0.2
Grade 5: District—4.8 | State—0.2
Grade 6: District—4.9 | State—0.2
Grade 7: District—4.8 | State—0.2
Grade 8: District—3.8 | State—0.2
Grade 10: District—4.0 | 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).

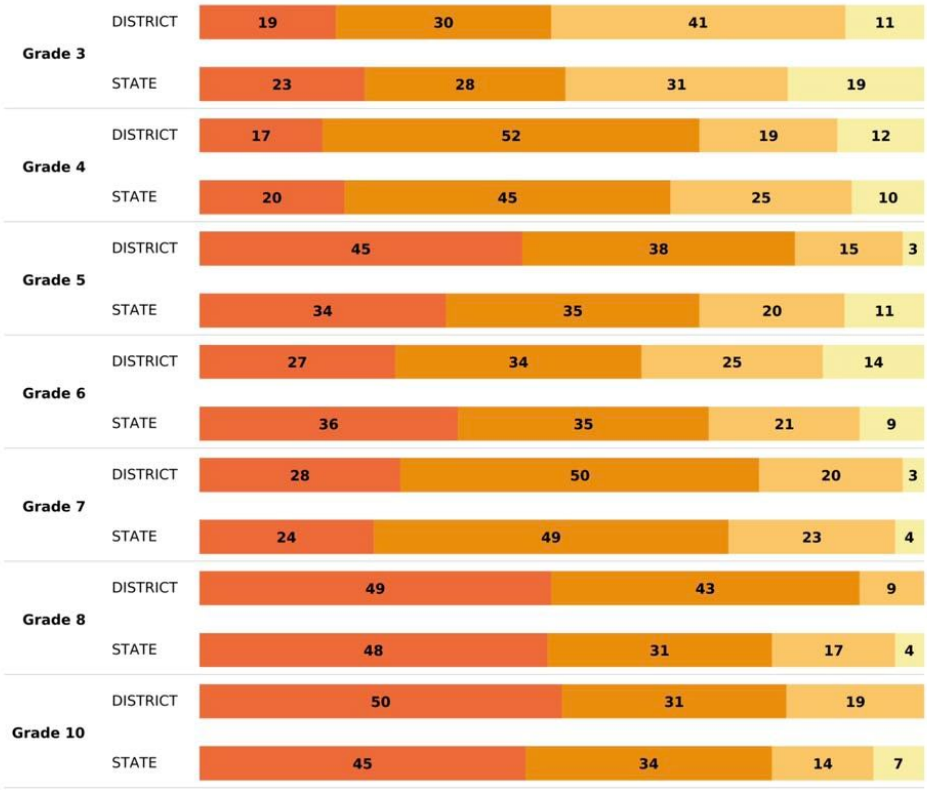
KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

DISTRICT REPORT SUBJECT: Mathematics

Percentage of Students in Each Performance Level, by Grade

Level 1 Level 2 Level 3 Level 4

Percentages may not add to 100% because of rounding.



KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

DISTRICT REPORT SUBJECT: Mathematics

Your District's Performance

(+) Exceeds (=) Meets (−) Below (×) Insufficient Data

Grade	3	4	5	6	7	8	10
SKILLS AND CONCEPTS	=	−	−	−	−	−	−
Operations and Algebraic Thinking	=	−					
Number and Operations in Base Ten		−	−				
Number and Operations with Fractions	=	−	−				
Measurement and Data	=	−	−				
Ratios and Proportional Relationships				=	−		
The Number System				−	−		
Expressions and Equations				−	−	−	
Algebra							−
Functions						−	−
Geometry	−			−	−	−	−
Statistics and Probability				−	−		−
STRATEGIC THINKING AND REASONING	−	−	−	−	−	−	−

SKILLS AND CONCEPTS
These questions require students to apply mathematical skills and concepts and interpret and carry out mathematical procedures with precision and fluency.

Operations and Algebraic Thinking
These questions require students to represent and solve problems with addition, subtraction, multiplication, and division; perform these operations with multidigit numbers; and explain patterns.

Number and Operations in Base Ten
These questions require students to demonstrate their understanding of place value by solving problems with multidigit numbers and decimals.

Number and Operations with Fractions
These questions require students to demonstrate their understanding that fractions represent parts of a whole, recognize that fractions can be written as decimals, and solve problems with fractions by applying their knowledge about working with whole numbers and decimals.

Measurement and Data
These questions require students to calculate time, volume, perimeter, area, and mass; measure angle size; convert measurements within a measurement system; represent and interpret measurement data; and use measurement skills to solve real-world problems.

Ratios and Proportional Relationships
These questions require students to use ratio reasoning and analyze proportional relationships to solve real-world and mathematical problems.

The Number System
These questions require students to divide fractions, find common factors and multiples, and perform operations with rational numbers.

KU ACHIEVEMENT & ASSESSMENT INSTITUTE
The University of Kansas

Expressions and Equations
These questions require students to solve equations that have variables and exponents, analyze relationships between dependent and independent variables and between proportional relationships, and use equations to model relationships and solve real-world problems.

Algebra
These questions require students to solve complex equations; construct, interpret, and graph equations that model data and represent relationships; and use equations to solve real-world problems.

Functions
These questions require students to interpret, compare, and build functions to model real-world relationships.

Geometry
These questions require students to describe the features of geometric figures, compare figures, apply geometric theorems, and solve real-world problems by applying formulas to figures.

Statistics and Probability
These questions require students to compare and draw inferences from data sets and to calculate probability of simple and compound events.

STRATEGIC THINKING AND REASONING
These questions require students to solve complex problems using problem-solving strategies and mathematical tools; explain their reasoning, defend their answers, and critique the reasoning of others; and analyze complex, real-world situations to construct and use mathematical models to solve problems, and to interpret results in the context of a situation.

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

Prediction on ACT scores is not available for mathematics grade 10 in 2022.

To learn more about the Kansas Assessment Program and these score reports, visit the "For Families" page on **ksassessments.org**.
For information on the Kansas Standards, visit **ksde.org**.

Kansans CAN